NON-SMOOTH, PARTICULARLY COMPOSITE, MULTI-OBJECTIVE OPTIMIZATION

Theory and algorithms

HIROKI TANABE

NON-SMOOTH, PARTICULARLY COMPOSITE, MULTI-OBJECTIVE OPTIMIZATION

Theory and algorithms

by

HIROKI TANABE

Submitted in partial fullfillment of
the requirement for the degree of
DOCTOR OF INFORMATICS
(Applied Mathematics and Physics)

KYOTO UNIVERSITY
KYOTO 606-8501, JAPAN
SEPTEMBER 2022

Preface

To accurately answer all human needs with optimization problems, we cannot avoid considering multi-objective optimization. Humankind is a *greedy* creature that cannot tolerate only a single desire and always has multiple preferences. Unfortunately, many of them conflict, and the best choice to answer all of them seldom exists. This trade-off is what makes multi-objective optimization a tough challenge. Even for problems indeed with multi-objectives, the single-objective models tend to be adopted. However, thanks to the long-standing wisdom of scientists, the development of theories and algorithms for multi-objective optimization has been gradually gaining speed in recent years. As one of the "*dwarfs who ride above the giants*," I would like to contribute to their development, even if only slightly.

This thesis provides theories and algorithms for problems in multi-objective optimization where the objective function is non-smooth. These types of problems are very complex. Thus, it is not practical to consider general non-smooth models for large-scale problems, which have been recently in high demand. Therefore, this thesis mainly focuses on multi-objective optimization problems with a specific structure, called composite models. In detail, this model's every objective function is the sum of differentiable and convex functions. Such models work well, for example, with the loss and regularization models in machine learning.

There are three main contributions of this thesis. One is new merit functions for multi-objective optimization and the elucidation of their properties. A merit function is a function that returns zero in the solution of the problem and a positive number otherwise. We can use it to reformulate the original problem and estimate the rate of convergence of the algorithm. Another contribution is the proximal gradient method for multi-objective optimization problems. It is a first-order method using information from first-order derivatives for composite multi-objective optimization problems. It is more efficient than existing first-order methods for non-smooth multi-objective problems; it has an $O(1/k)$ convergence rate. It can

also generate stationary points for non-convex problems. Another contribution is the accelerated proximal gradient method for multi-objective optimization. It does not work for non-convex problems but is faster than the proximal gradient method and solves problems with $O(1/k^2)$. The proposed algorithm is also novel for single-objective problems if the parameters are well-chosen. Its numerical results are better than existing algorithms for single objectives.

Hiroki Tanabe
September 2022

Acknowledgment

This thesis summarizes the author's research during the enrollment in the doctoral course at the Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University. Professor Nobuo Yamashita and Associate Professor Ellen Hidemi Fukuda of the same department, my supervisors, allowed me to conduct this research and provided guidance throughout it. I want to express my deepest gratitude to them.

I likewise thank the members of the System Optimization Laboratory, the scientists of the Operations Research Society of Japan, and many others for their valuable comments and suggestions at the workshops and conferences. I would also like to thank my friends and family for their emotional support.

Part of this research was supported by Grant-in-Aid for JSPS Fellows (20J21961) from the Japan Society for the Promotion of Science.

Contents

List of Figures

List of Tables

List of Symbols and Notations

(x, y) the open line segment between x and y

$[x, y]$ the closed line segment between x and y

$\mathrm{conv}(C)$ the convex hull of C

$\dim(X)$ the dimension of a space X

$\mathrm{dist}(x, C)$ the distance between x and C

$\mathrm{dom}(f)$ the effective domain of function f

$\langle x, y \rangle$ the Euclidean inner product between x and y

$\mathrm{int}(C)$ the interior of C

$\mathcal{J}_f(x)$ the Jacobian matrix of f at x

$\ker(A)$ the kernel of a matrix A

$\nabla f(x)$ the gradient of f at x

$\|x\|_1$ the ℓ_1-norm of x

$\|x\|_2$ the ℓ_2-norm of x

$\|x\|_\infty$ the ℓ_∞-norm of x

$\partial f(x)$ the subdifferential of f at x

$\mathbf{R}$ the set of real numbers

$\mathbf{R}^n$ the n-dimensional real space

$\mathbf{R}^n_+$ the nonnegative orthant in $\mathbf{R}^n$

Δ^n	the unit n-simplex
I_n	the $n \times n$ identity matrix

Chapter 1

Introduction

Optimization, a branch of applied mathematics, minimizes (or maximizes) an objective function under given constraints. It is a fundamental technique for operations research and machine learning.

This chapter first describes multi-objective optimization, the subject of this thesis, and composite optimization, a crucial class of non-smooth optimization. It also explains the merit function, an analytical tool for optimization. Finally, it identifies the research challenges on multi-objective optimization problems and explains this thesis's motivations, contributions, and outlines.

1.1 Multi-objective optimization

Optimization problems usually deal with only one objective function. However, many real-world problems have multiple objectives. One solution to this is *multi-objective optimization*, which minimizes several objective functions as follows:

$$\min_{x \in C} \quad F(x), \tag{1.1}$$

where $C \subseteq \mathbf{R}^n$ is a constraint set, and $F \colon \mathbf{R}^n \to (-\infty, +\infty]^m$ is a vector-valued function with $F \coloneqq (F_1, \ldots, F_m)^\top$. When $m = 1$, (1.1) reduces to a single-objective optimization. This model has many applications in engineering [**Eschenauer1990**], statistics [**Carrizosa1998**], and machine learning (particularly multi-task learning [**Sener2018**, **Lin2019**] and neural architecture search [**Kim2017**, **Dong2018**, **Elsken2019**]).

In most cases of $m \geq 2$, no single point minimizes all objective functions simul-

taneously, so we use the concept of *Pareto optimality*, a generalization of the usual optimality for single-objective problems. We say that $y \in C$ *Pareto dominates* $x \in C$ if $F_i(y) \leq F_i(x)$ for all $i = 1, \ldots, m$ and $F_j(y) < F_j(x)$ for at least one $j = 1, \ldots, m$, and we call a point *Pareto optimal* if it is not Pareto dominated by any other point. Generally, the Pareto optimal solution is not unique and constitutes a set. We call such a set the *Pareto frontier*. The solutions in the Pareto frontier are in trade-off relationships, and the decision-makers must select a solution from it further.

1.1.1 Scalarization approach

The *scalarization approach* [**Gass1955**, **Geoffrion1968**, **Zadeh1963**] is one of the most popular strategies for multi-objective problems. It converts the original multi-objective problem into a parameterized scalar-valued problem.

Let us now introduce the *weighted sum method* [**Zadeh1963**], one of the most well-known scalarization techniques. It scalarizes (1.1) with the weight vector $w := (w_1, \ldots, w_m)^\top \in \mathbf{R}^m$ as follows:

$$\min_{x \in \mathbf{R}^n} \quad \langle w, F(x) \rangle, \tag{1.2}$$

where

$$w \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} w_i = 1.$$

When F is convex, for every Pareto optimal solution x^* of (1.1), there exists w such that x^* is the solution of (1.2) [**Miettinen1998**]. However, it may be challenging to choose a *good* weight in advance. Moreover, if F is non-convex, there may be Pareto optimal solutions that are not the solutions of (1.2) for any w, and some w may make (1.2) unbounded.

1.1.2 Heuristics

Heuristics are approaches that do not necessarily lead to the optimal solution but can yield a solution close to the optima at some level. Regarding the multi-objective context, in many cases, heuristics employ evolutionary algorithms, particularly genetic algorithms (GA) such as NSGA-II [**Deb2002**] and NSGA-III [**Deb2014**], being practical for the Pareto frontier enumeration because they are multi-point search algorithms. These approaches have had some success for real-world problems, but

they have the disadvantage that there is no theoretical convergence guarantee to obtain a Pareto solution.

1.1.3 Descent methods

Descent methods [**Fukuda2014**] are iterative algorithms that decrease the objective function values at each iteration. They do not require *a priori* parameters selection like scalarization, and unlike heuristics, we can analyze their global convergence property under reasonable assumptions. All algorithms proposed in this thesis are part of the descent methods. Below we provide typical descent methods for (1.1).

Example 1.1
The steepest descent method [Fliege2000]

Consider a smooth unconstrained multi-objective optimization, i.e., $C = \mathbf{R}^n$ and each F_i is differentiable in (1.1). Then, the steepest descent method updates $\{x^k\}$ by the following operations:

$$\begin{aligned} d^k &:= \operatorname*{argmin}_{d\in\mathbf{R}^n}\left[\max_{i=1,\dots,m}\left\langle \nabla F_i(x^k), d\right\rangle + \frac{1}{2\alpha_k}\|d\|_2^2\right], \\ x^{k+1} &:= x^k + s_k d^k \end{aligned} \tag{1.3}$$

*with $\alpha_k > 0$ and $s_k > 0$. When $m = 1$, we have $d^k = -\alpha_k \nabla F_1(x)$, which is the steepest descent direction for the scalar optimization [**Cauchy1847**].*

The projected gradient method [Grana-Drummond2004, Fukuda2013]

For a convex-constrained smooth multi-objective optimization, i.e., $C \subseteq \mathbf{R}^n$ is non-empty, closed, and convex, and every F_i is differentiable in (1.1), we can use the projected gradient method described by

$$\begin{aligned} z^k &:= \operatorname*{argmin}_{z\in C}\left[\max_{i=1,\dots,m}\left\langle \nabla F_i(x^k), z - x^k\right\rangle + \frac{1}{2\alpha_k}\left\|z - x^k\right\|_2^2\right], \\ x^{k+1} &:= x^k + s_k\left(z^k - x^k\right) \end{aligned} \tag{1.4}$$

*with $\alpha_k > 0$ and $s_k > 0$. When $m = 1$, (1.4) reduces to the projected gradient method for scalar optimization [**Polyak1963**, **Goldstein1964**,*

Goldstein1967, **McCormick1969**], *i.e.*,

$$z^k := \mathbf{proj}_C(x^k - \alpha_k \nabla F_1(x^k)),$$
$$x^{k+1} := x^k + s_k(z^k - x^k),$$

where $\mathbf{proj}_C$ *denotes the projection onto* C *given by*

$$\mathbf{proj}_C(x) := \operatorname*{argmin}_{z \in C} \|z - x\|_2. \tag{1.5}$$

Moreover, when $C = \mathbf{R}^n$, (1.4) *amounts to the steepest descent method* (1.3).

The projected subgradient method [Bello-Cruz2013]

Focus on a convex-constrained, non-smooth, and convex multi-objective optimization, i.e., C *is a non-empty, closed, and convex subset of* $\mathbf{R}^n$, *and each* F_i *is convex and non-differentiable in* (1.1). *The subgradient method requires an exogenous sequence* $\{\beta_k\}$ *satisfying*

$$\beta_k > 0, \quad \sum_{k=0}^{\infty} \beta_k = \infty, \quad \textit{and} \quad \sum_{k=0}^{\infty} \beta_k^2 < \infty$$

and generates $\{x^k\}$ *by*

$$x^{k+1} := \operatorname*{argmin}_{z \in C} \left[\frac{1}{2} \|z - x^k\|_2^2 + \frac{\beta_k}{\eta_k} \max_{i=1,\ldots,m} \langle \xi_i^k, z - x^k \rangle \right],$$

where $\xi_i^k \in \partial F_i(x^k)$ *and*

$$\eta_k := \max_{i=1,\ldots,m} \|\xi_i\|.$$

When $m = 1$, *this step represents the projected subgradient method [***Polyak1967**, **Polyak1969**, **Shor1985**, **Alber1998**, **Alber2001***] for scalar optimization:*

$$x^{k+1} := \mathbf{proj}_C\left(x^k - \frac{\beta_k}{\eta_k} \xi_1^k \right).$$

1.2 Composite optimization

Composite optimization has the following structure:

$$\min_{x\in\mathbf{R}^n} \quad F(x) \coloneqq f(x) + g(x), \tag{1.6}$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ is L_f-smooth with some $L_f > 0$, and $g\colon \mathbf{R}^n \to (-\infty, +\infty]$ is closed, proper, and convex. When f is convex, we call (1.6) *convex composite*. This model has many applications, particularly in machine learning. In detail, f and g often represent the loss function and the regularization term, respectively. We list below some typical examples with the structure (1.6).

Example 1.2

Smooth unconstrained minimization

If $g = 0$, (1.6) reduces to the unconstrained smooth minimization

$$\min_{x\in\mathbf{R}^n} \quad f(x),$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ is L_f-smooth.

Convex-constrained smooth minimization

If g is an indicator function of a non-empty, closed, and convex set C, i.e.,

$$g(x) = \delta_C(x) \coloneqq \begin{cases} 0 & x \in C, \\ \infty & \text{otherwise}, \end{cases} \tag{1.7}$$

then (1.6) amounts to the convex-constrained smooth minimization

$$\min_{x\in C} \quad f(x)$$

with an L-smooth function f.

ℓ_1-regularization

If $g(x) \coloneqq \tau\|x\|_1$ for some $\tau > 0$, (1.6) reduces to the ℓ_1-regularizaiton

$$\min_{x\in C} \quad f(x) + \tau\|x\|_1$$

with f being L_f-smooth.

1.2.1 The proximal gradient method

The *proximal gradient method* [**Fukushima1981**] is one of the most common algorithms for solving (1.6). For a given $x^0 \in \text{int}(\text{dom}(F))$, it recursively update $\{x^k\}$ by

$$x^{k+1} = \mathbf{prox}_{\alpha_k g}\big(x^k - \alpha_k \nabla f\big(x^k\big)\big),$$

where **prox** is the *proximal operator*, which we will define in (2.10). If we can estimate the Lipschitz constant L_f, we can use a constant stepsize $\alpha_k \in (0, 1/L_f]$. Otherwise, we can determine α_k in each iteration by backtracking.

The description of the algorithm now follows.

Algorithm 1.1 The proximal gradient method

Input: $x^0 \in \text{int}(\text{dom}(F)), \varepsilon > 0$

1: $k \leftarrow 0$
2: **repeat**
3: pick $\alpha_k > 0$
4: $x^{k+1} \leftarrow \mathbf{prox}_{\alpha_k g}(x^k - \alpha_k \nabla f\big(x^k\big))$
5: $k \leftarrow k + 1$
6: **until** $\|x^k - x^{k-1}\|_\infty < \varepsilon$
7: **return** x^k

With this algorithm, $\{\|x^{k+1} - x^k\|_2\}$ converges to zero with a rate of $O\Big(\sqrt{1/k}\Big)$ and every accumulation point of $\{x^k\}$, if it exists, is a stationary point [**Beck2017**]. When f is convex, $\{x^k\}$ converges to the global minima x^*, and $\{F\big(x^k\big) - F(x^*)\}$ converges to zero with a rate of $O(1/k)$ [**Beck2017**]. Moreover, when f is strongly convex, $\{x^k\}$ converges linearly to x^* [**Beck2017**]. Furthermore, if we assume the so-called proximal-PL condition, which we will define by (2.14), $\{F\big(x^k\big)\}$ converges linearly to $F(x^*)$ [**Karimi2016**].

1.2.2 The accelerated proximal gradient method

When f is convex, the accelerated proximal gradient method, also known as the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [**Beck2009**], can solve (1.6) with an $O(1/k^2)$ rate of convergence, while the proximal gradient method achieves a rate of $O(1/k)$.

We describe below the algorithm. Like the proximal gradient method, the stepsize α_k may be fixed at a constant or updated by the backtracking procedure.

Algorithm 1.2 The accelerated proximal gradient method

Input: $x^0 \in \text{int}(\text{dom}(F)), \varepsilon > 0$
1: $k \leftarrow 1$
2: $y^1 \leftarrow x^0$
3: $t_1 \leftarrow 1$
4: **repeat**
5: pick $\alpha_k > 0$
6: $x^k \leftarrow \mathbf{prox}_{\alpha_k g}(y^k - \alpha_k \nabla f(y^k))$
7: $t_{k+1} \leftarrow \sqrt{t_k^2 + 1/4} + 1/2$
8: $\gamma_k \leftarrow (t_k - 1)/t_{k+1}$
9: $y^{k+1} \leftarrow x^k + \gamma_k(x^k - x^{k-1})$
10: $k \leftarrow k + 1$
11: **until** $\|x^k - y^k\|_\infty < \varepsilon$
12: **return** x^k

With Algorithm 1.2, $\{F(x^k) - F(x^*)\}$ for the global minima x^* converges to zero with a rate of $O(1/k^2)$ [**Beck2009**], but the convergence of iterates remains unknown. With a slite modification, that is, changing the update rule of the momentum factor with $t_k = (k+a-1)/a$ for some $a > 2$, we can prove that $\{x^k\}$ converges to the minima while keeping the convergence rate of $O(1/k^2)$ [**Chambolle2015**].

1.3 Merit functions

Merit functions [**Fukushima1996**] are maps that return zeros at the problems' solutions and strictly positive values otherwise. In other words, they are the objective functions of optimization problems with the same solutions as the original problems. Therefore, the merit functions should have the following properties:

- Quick computability;
- Continuity;
- Differentiability;
- Optimality of the stationary points;
- Level-boundedness;
- Error-boundedness.

Moreover, as we can consider the merit functions to represent how far feasible points are from the optimal solutions, they help analyze convergence rates of the optimization algorithm.

1.3.1 Merit functions for variational inequalities

Merit functions have evolved in the context of reformulating variational inequalities (VIs) and complementarity problems (CPs) as optimization problems [**Fukushima1996**]. The *variational inequality* (VI) consists in finding $x \in C$ such that

$$\langle T(x), y - x\rangle \geq 0 \quad \text{for all } y \in C, \tag{1.8}$$

where $C \subseteq \mathbf{R}^n$ is nonempty, closed, and convex, and $T\colon \mathbf{R}^n \to \mathbf{R}^n$ is continuous. We can also rewrite (1.8) as the following *complementarity problem* (CP):

$$T(x) \geq 0, \quad x \geq 0, \quad \text{and} \quad \langle T(x), x\rangle \geq 0. \tag{1.9}$$

In particular, if T is affine, we call (1.9) the *linear complementarity problem* (LCP). There are many merit functions for VIs and CPs, but here we illustrate the most basic two merit functions for VIs.

Example 1.3 (Merit functions for the VI (1.8))
The classical gap function [Auslender1976, Hearn1982]

We call the function $G_\infty\colon \mathbf{R}^m \to (-\infty, +\infty]$ *the* classical gap function*:*

$$G_\infty(x) := \sup_{y \in C} \langle T(x), x - y\rangle. \tag{1.10}$$

It has the following properties:

- $G_\infty(x) \geq 0$ *for all* $x \in C$*;*
- $G_\infty(x) = 0$ *and* $x \in C$ *if and only if* x *satisfies* (1.8)*;*
- *If* C *is bounded,* G_∞ *is finite everywhere.*

The top two indicate that G_∞ *is a merit function for the VI* (1.8)*.*

The regularized gap function [Fukushima1992, Auchmuty1989]

For a given parameter $\alpha > 0$*, we can consider the* regularized gap func-

tion $G_\alpha\colon \mathbf{R}^n \to \mathbf{R}$ *defined by*

$$G_\alpha(x) \coloneqq \max_{y\in C}\left[\langle T(x), x-y\rangle - \frac{1}{2\alpha}\|x-y\|_2^2\right], \tag{1.11}$$

which is a merit function for the VI (1.8), too. Since (1.11) maximizes a strongly concave function on a nonempty, closed, and convex set, even if C is unbounded, a unique point attains the maximum, and G_α is finite everywhere. Moreover, denoting such a maximizer by $H_\alpha(x)$, if T is continuously differentiable, G_α is also differentiable at any point x, and we have

$$\nabla G_\alpha(x) = T(x) - [\mathcal{J}_T(x) - \alpha^{-1}I_n](H_\alpha(x) - x).$$

Note that

$$H_\alpha(x) = \mathbf{proj}_C(x - \alpha T(x)).$$

Furthermore, if the Jacobian $\mathcal{J}_T(x)$ is positive definite on C, any stationary point of the problem

$$\min_{x\in C} \quad G_\alpha(x)$$

*solves the VI (1.8) [**Fukushima1992**]. In addition, if T is strongly monotone with modulus $\mu > 0$, i.e.,*

$$\langle T(x) - T(x'), x - x'\rangle > \mu\|x - x'\|^2 \quad \text{for all } x, x' \in \mathbf{R}^n,$$

*and if $\alpha > 1/(2\mu)$, then G_α has the following error bound property [**Taji1993**]:*

$$\|x - x^*\| \le \sqrt{\frac{G_\alpha(x)}{\mu - 1/(2\alpha)}} \quad \text{for all } x \in S,$$

where x^* is the unique solution of the VI (1.8).

1.3.2 Merit functions for multi-objective problems

The history of research on merit functions for multi-objective problems is relatively new, beginning in 1998 with **Chen1998empty citation** on (1.1) under the assumptions of polyhedrality of C and convexity of F. Afterward, various merit functions appeared for multi-objective problems, including multi-objective optimization [**Liu2009**, **Dutta2017**], (finite-dimensional) vector variational inequali-

ties [**Chen2000**, **Konnov2005**, **Li2005**, **Yang2002**, **Yang2003**, **Charitha2010**, **Li2010**], and (finite-dimensional) vector equilibrium problems [**Huang2007**, **Li2005**, **Li2007**, **Li2006**, **Mastroeni2003**]. Below we pick up generalizations of Example 1.3 to the *weak Stamnpacchia type vector variational inequality* $(SVVI)^w$, which consists in finding $x \in C$ such that

$$(\langle T_1(x), y - x\rangle, \ldots, \langle T_m(x), y - x\rangle) \notin -\operatorname{int}(\mathbf{R}^m_+) \quad \text{for all } y \in C, \tag{1.12}$$

where $C \subseteq \mathbf{R}^n$ is a nonempty, closed, convex, and $T_i \colon \mathbf{R}^n \to \mathbf{R}^n, i = 1, \ldots, m$. Note that x satisfies (1.12) if and only if x is weakly Pareto optimal for (1.1) when F_i is differentiable and $T_i = \nabla F_i$ for each $i = 1, \ldots, m$.

Example 1.4
The gap function for $(SVVI)^w$ [Charitha2010, Li2010]

We can write the gap function $G_\infty \colon \mathbf{R}^n \to (-\infty, +\infty]$ for (1.12) as

$$G_\infty(x) := \min_{\lambda \in \Delta^m} \sup_{y \in C} \left\langle \sum_{i=1}^m \lambda_i T_i(x), x - y \right\rangle,$$

where Δ^m is the unit m-simplex, which we will define by (2.1). When $m = 1$, it corresponds to (1.10). Like (1.10), G_∞ is a merit function for (1.12), i.e.,

- *$G_\infty(x) \geq 0$ for all $x \in C$;*
- *$G_\infty(x) = 0, x \in C$ if and only if x solves (1.12),*

and it is finite-valued if C is bounded.

The regularized gap function for $(SVVI)^w$ [Charitha2010]

We can define the regularized gap function $G_\alpha \colon \mathbf{R}^n \to \mathbf{R}$ with $\alpha > 0$ for (1.12) by

$$G_\alpha(x) := \min_{\lambda \in \Delta^m} \max_{y \in C} \left[\left\langle \sum_{i=1}^m \lambda_i T_i(x), x - y \right\rangle - \frac{1}{2\alpha} \|x - y\|_2^2 \right], \tag{1.13}$$

matching (1.11) when $m = 1$. It also satisfies the two propeties as a merit function for (1.12). Moreover, if each $T_i, i = 1, \ldots, m$ is continuously differentiable, then G_α is directionally differentiable in any direction $d \in \mathbf{R}^n$,

and

$$G'_\alpha(x;d) = \min_{\lambda\in\Lambda(x)} \left[\left\langle \sum_{i=1}^m \lambda_i T_i(x) - \sum_{i=1}^m \lambda_i \mathcal{J}_{T_i}(x)(H_\alpha(x,\lambda)-x), d \right\rangle + \alpha\langle H_\alpha(x,\lambda)-x, d\rangle \right],$$

where

$$\begin{aligned} H_\alpha(x,\lambda) &:= \mathbf{proj}_C\left(x - \alpha^{-1}\sum_{i=1}^m \lambda_i T_i(x)\right), \\ T_\alpha(x,\lambda) &:= -\left\langle \sum_{i=1}^m \lambda_i T_i(x), H_\alpha(x,\lambda)-x \right\rangle - \frac{1}{2\alpha}\|H_\alpha(x,\lambda)-x\|_2^2, \\ \Lambda(x) &:= \{\lambda\in\Delta^m \mid G_\alpha(x) = T_\alpha(x,\lambda)\}. \end{aligned}$$

Particularly, if $\Lambda(x)$ is a singleton, i.e., $\Lambda(x)=\{\lambda(x)\}$, G_α is Gateaux differentiable at x and

$$\begin{aligned} &\nabla G_\alpha(x) \\ &= \sum_{i=1}^m \lambda_i(x) T_i(x) - \sum_{i=1}^m \lambda_i(x)\mathcal{J}_{T_i}(x)[H_\alpha(x,\lambda(x))-x] + \alpha^{-1}[H_\alpha(x,\lambda(x))-x]. \end{aligned}$$

Furthermore, if each $T_i, i=1,\ldots,m$ is strongly monotone with modulus $\mu_i>0$, and if $\alpha > 1/(2\mu)$ with $\mu := \min_{i=1,\ldots,m}\mu_i$, then G_α provides the error bound:

$$\mathrm{dist}(x, sol(SVVI)^w) \le \sqrt{\frac{G_\alpha(x)}{\mu - 1/(2\alpha)}} \quad \textit{for all } x\in C,$$

where $sol(SVVI)^w$ denotes the solution set of (1.12).

1.4 Motivations and contributions

As discussed in Section 1.1, multi-objective optimization (1.1) is an indispensable model in dealing with real-world problems, and the studies on its theories and algorithms have great significance. On the other hand, many previous studies on

multi-objective optimization, particularly on the descent methods described in Section 1.1.3 and the merit functions described in Section 1.3.2, have dealt with smooth problems, and there is still room for exploration of non-smooth problems. The projected subgradient method introduced in Example 1.1 can handle non-smooth multi-objective optimization, but it may not work well for large-scale problems due to the stepsize decay.

This thesis focuses on non-smooth multi-objective optimization problems with specific structures, mainly the generalization of the composite model introduced in Section 1.2, i.e.,

$$\min_{x \in C} \quad F(x) = f(x) + g(x), \tag{1.14}$$

where $C \subseteq \mathbf{R}^n$ is a non-empty, closed, and convex set, and $F\colon \mathbf{R}^n \to (-\infty, +\infty]^m, f\colon \mathbf{R}^n \to \mathbf{R}^m, g\colon \mathbf{R}^n \to (-\infty, +\infty]^m$ are vector-valued functions with $F \coloneqq (F_1, \ldots, F_m)^\top, f \coloneqq (f_1, \ldots, f_m)^\top, g \coloneqq (g_1, \ldots, g_m)^\top$ such that $f_i\colon \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable and $g_i\colon \mathbf{R}^n \to (-\infty, +\infty]$ is closed, proper, and convex. Then, we presents their theory and algorithms.

1.5 Outline of the thesis

After introducing in Chapter 2 some symbols, basic definitions, and their properties necessary for the discussion, Chapter 3 proposes and characterizes three new types of merit functions for non-smooth multi-objective optimization problems: the gap function for continuous problems, the regularized gap function for convex problems, and the regularized and partially linearized gap functions for composite problems. Chapter 4 develops the proximal gradient method for composite multi-objective optimization problems, describes its convergence, convergence rate, applications to robust multi-objective optimization, and performs numerical experiments. Chapter 5 presents its acceleration applicable with *convex* composite objectives: the accelerated proximal gradient method or Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) and provides similar discussions. We note here that our multi-objective FISTA represents a new algorithm even for single objectives, depending on the choice of acceleration factors, and performs better in numerical experiments.

Chapter 2

Preliminaries

This chapter presents some notations, basic definitions, and their properties used in this thesis.

2.1 Vectors and matrices

Let $\mathbf{R}^p$ denote the space of p-dimensional real column vectors. Meanwhile, we write the set of real numbers as simply $\mathbf{R}$ or $(-\infty, +\infty)$ instead of $\mathbf{R}^1$. Moreover, $\mathbf{R}^{q\times p}$ stands for the space formed by $q \times p$ real matrices. In addition, define the nonnegative orthant in $\mathbf{R}^p$ by

$$\mathbf{R}^p_+ \coloneqq \{v \in \mathbf{R}^p \mid v_i \geq 0, i = 1, \ldots, p\},$$

and define the unit simplex $\Delta^p \subseteq \mathbf{R}^p_+$ by

$$\Delta^p \coloneqq \left\{ v \in \mathbf{R}^p_+ \;\middle|\; \sum_{i=1}^{p} v_i = 1 \right\}. \tag{2.1}$$

The orthant $\mathbf{R}^p_+$ induces the partial orders for any $v^1, v^2 \in \mathbf{R}^p$: $v^1 \leq v^2$ (alternatively, $v^2 \geq v^1$ if $v^2 - v^1 \in \mathbf{R}^p_+$, and $v^1 < v^2$ (alternatively, $v^2 > v^1$) if $v^2 - v^1 \in \mathrm{int}(\mathbf{R}^p_+)$. In other words, we say that $v^1 \leq (<)\ v^2$ if $v^1_i \leq (<)\ v^2_i$ for all $i = 1, \ldots, p$. Furthermore, let $\langle \cdot, \cdot \rangle$ stand for the Euclidean inner product, i.e., $\langle v^1, v^2 \rangle \coloneqq \sum_{i=1}^p v^1_i v^2_i$. We also define ℓ_2-norm $\|\cdot\|_2$, ℓ_1-norm $\|\cdot\|_1$, and ℓ_∞-

norm $\|\cdot\|_\infty$ by

$$\|v\|_2 \coloneqq \sqrt{\langle v, v\rangle} = \sum_{i=1}^{p} v_i^2, \quad \|v\|_1 \coloneqq \sum_{i=1}^{p} |v_i|, \quad \text{and} \quad \|v\|_\infty \coloneqq \max_{i=1,\dots,p} |v_i|$$

for any $v \in \mathbf{R}^p$. Finally, we note some apparent inequalities that hold for arbitrary $v^1, v^2, v^3 \in \mathbf{R}^p$ and $\{v^{s,p}\} \subseteq \mathbf{R}^p$.

$$\min_{i=1,\dots,p} v_i^1 - \min_{i=1,\dots,p} v_i^2 \geq \min_{i=1,\dots,p} \bigl(v_i^1 - v_i^2\bigr), \tag{2.2}$$

$$\bigl\|v^2 - v^1\bigr\|^2 + 2\bigl\langle v^2 - v^1, v^1 - v^3\bigr\rangle = \bigl\|v^2 - v^3\bigr\|^2 - \bigl\|v^1 - v^3\bigr\|^2, \tag{2.3}$$

$$\sum_{s=1}^{r}\sum_{p=1}^{s} v^{s,p} = \sum_{p=1}^{r}\sum_{s=p}^{r} v^{s,p}. \tag{2.4}$$

2.2 Convexity and semi-continuity

We first define the convexity of sets and functions. A set $C \subseteq \mathbf{R}^p$ is *convex* if

$$(1-\alpha)v^1 + \alpha v^2 \in C \quad \text{for all } v^1, v^2 \in C, \alpha \in [0,1].$$

Likewise, a function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *convex* if

$$h((1-\alpha)x + \alpha y) \leq (1-\alpha)h(x) + \alpha h(y) \quad \text{for all } x, y \in \operatorname{dom}(h), \alpha \in [0,1],$$

strictly convex if

$$h((1-\alpha)x + \alpha y) < (1-\alpha)h(x) + \alpha h(y) \quad \text{for all } x, y \in \operatorname{dom}(h), \alpha \in (0,1),$$

and μ_f*-convex* with $\mu_f \in \mathbf{R}$ if

$$h((1-\alpha)x + \alpha y) \leq (1-\alpha)h(x) + \alpha h(y) - \frac{\mu_f}{2}\alpha(1-\alpha)\|x-y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom}(h), \alpha \in [0,1],$$

where $\operatorname{dom}(h)$ stands for the *effective domain* of h given by

$$\operatorname{dom}(h) \coloneqq \{x \in \mathbf{R}^p \mid h(x) < \infty\}.$$

In particular, the *strong convexity* (with modulus μ_f) denotes the μ_f-convexity with $\mu_f > 0$. We also note that the 0-convexity is equivalent to the usual con-

vexity. Moreover, if $\mathrm{dom}(h) \neq \emptyset$ for some convex function $h\colon (-\infty, +\infty]$, we say that h is *proper* and convex. On the other hand, we call h to be concave if $-h$ is convex. Every definition and argument relating to convex functions also holds for concave functions by appropriately interchanging $\leq$ and $\geq$, $+\infty$ and $-\infty$, sup and inf, etc.

Let us now introduce the semi-continuity of functions. For all $\{x^k\} \subseteq \mathbf{R}^p$ converging to $x \in \mathbf{R}^p$, a function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *upper semi-continuous* at x if

$$h(x) \geq \limsup_{k\to\infty} h\big(x^k\big)$$

and *lower semi-continuous* if

$$h(x) \leq \liminf_{k\to\infty} h\big(x^k\big).$$

A necessary and sufficient condition for h to be lower semi-continuous is that the level set $\mathbf{lev}_c(h)$ given by

$$\mathbf{lev}_c(h) \coloneqq \{x \in \mathbf{R}^p \mid h(x) \leq c\} \tag{2.5}$$

is closed for any $c \in \mathbf{R}$. We refer to lower-semi-continuous, proper, and convex functions as *closed, proper, and convex* functions. The level sets of convex functions are convex, and the level sets of closed, proper, and convex functions are closed and convex. Note that if $\mathbf{lev}_c(h)$ is bounded for all $c \in \mathbf{R}$, we say that h is *level-bounded*. For example, every strongly convex function is level-bounded. Note also that (2.5) is applicable as the definition of the level set for the vector-valued function $h\colon \mathbf{R}^p \to \mathbf{R}^q$ and $c \in \mathbf{R}^q$.

2.3 Differentiability

Suppose that $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is finite-valued in an appropriate neighborhood of $x \in \mathbf{R}^p$. If h has the partial derivative

$$\frac{\partial h(x)}{\partial x_i} \coloneqq \lim_{\alpha\to 0} \frac{h(x + \alpha e^i) - h(x)}{\alpha} \quad \text{for all } i = 1, \ldots, p$$

with e^i being the unit vector along the x_i-axis, and if

$$h(x+\varepsilon) = h(x) + \langle \nabla h(x), \varepsilon \rangle + o(\|\varepsilon\|_2) \quad \text{for all } \varepsilon \in \mathbf{R}^p \tag{2.6}$$

with $o\colon [0,+\infty) \to \mathbf{R}$ satisfying $\lim_{t\to 0} o(t)/t = 0$ and

$$\nabla h(x) \coloneqq \begin{bmatrix} \dfrac{\partial h(x)}{\partial x_1} \\ \vdots \\ \dfrac{\partial h(x)}{\partial x_p} \end{bmatrix},$$

then h is *differentiable* at x, and we call $\nabla h(x) \in \mathbf{R}^p$ a *gradient* of h at x. If $\nabla h(x)$ is continuous at x, we say that h is *continuously differentiable* at x. Again, if h has second-order derivatives and

$$h(x+\varepsilon) = h(x) + \langle \nabla h(x), \varepsilon \rangle + \frac{1}{2}\big\langle \varepsilon, \nabla^2 h(x)\varepsilon \big\rangle + o\big(\|h\|_2^2\big)$$

with

$$\nabla^2 h(x) \coloneqq \begin{bmatrix} \dfrac{\partial^2 h(x)}{\partial x_1\, \partial x_1} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_1\, \partial x_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 h(x)}{\partial x_p\, \partial x_1} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_p\, \partial x_p} \end{bmatrix},$$

then h is *twice differentiable* at $x \in \mathbf{R}^p$, and $\nabla^2 h(x)$ is a *Hessian matrix* of h at x. When $\nabla^2 h$ is continuous at x, h is *twice continuously differentiable* at x, and then $\nabla^2 h(x)$ is symmetric. On the other hand, for a vector-valued function $h\colon \mathbf{R}^p \to \mathbf{R}^q$ with $h \coloneqq (h_1, \ldots, h_m)^\top$, $\mathcal{J}_h(x)$ denotes the Jacobian matrix of h at x, that is,

$$\mathcal{J}_h(x) \coloneqq \begin{bmatrix} \dfrac{\partial h_1(x)}{\partial x_1} & \cdots & \dfrac{\partial h_1(x)}{\partial x_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_q(x)}{\partial x_1} & \cdots & \dfrac{\partial h_q(x)}{\partial x_p} \end{bmatrix} = [\nabla h_1(x), \ldots, \nabla h_q(x)]^\top \in \mathbf{R}^{q\times p}, \tag{2.7}$$

where $\top$ denotes transpose.

2.4 Hölder and Lipschitz continuity

We call $h\colon \mathbf{R}^p \to \mathbf{R}$ to be *locally Hölder continuous* with exponent $\beta > 0$ if for every bounded set $\Omega \subseteq \mathbf{R}^p$ there exists $L_h > 0$ such that

$$|h(x) - h(y)| \leq L_h \|x - y\|_2^\beta \quad \text{for all } x, y \in \Omega.$$

In particular, when L_h does not depend on Ω, we say that h is Hölder continuous with exponent $\beta > 0$. Moreover, we refer to the (local) Hölder continuity with exponent 1 as the *(local) Lipschitz continuity*. When h is Lipschitz continuous, we call L_h the *Lipschitz constant*, and we also say that h is L_h*-Lipschitz continuous*. As the following lemma shows, many functions with *good* properties are locally Lipschitz continuous.

Lemma 2.1
Continuously differentiable functions and finite-valued convex functions are locally Lipschitz continuous.

Proof. The former is due to the mean value theorem, and the latter is from [**WayneStateUniversity1972**]. □

Furthermore, if h is continuously differentiable and ∇h is L_h-Lipschitz continuous, we say that h is L_h-smooth. We now recall the so-called descent lemma [**Bertsekas1999**] as follows:

Lemma 2.2 (Descent Lemma [Bertsekas1999])
Let $h\colon \mathbf{R}^p \to \mathbf{R}$ is L_h-smooth on $\mathbf{R}^p$ with $L_h > 0$. Then, we have

$$|h(y) - h(x) - \langle \nabla h(x), y - x \rangle| \leq \frac{L_h}{2} \|x - y\|_2^2 \quad \textit{for all } x, y \in \mathbf{R}^p.$$

2.5 Directional derivatives and subgradients

A function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is *directionally differentiable* at $x \in \mathbf{R}^p$ in a direction $d \in \mathbf{R}^p$ if

$$h'(x; d) \coloneqq \lim_{\beta \searrow 0} \frac{h(x + \beta d) - h(x)}{\beta} \tag{2.8}$$

exists, and then we call $h'(x; d)$ the *directional derivative* at x in a direction d. When h is differentiable at x, we have $h'(x; d) = \langle \nabla h(x), d \rangle$ for all $d \in \mathbf{R}^p$. As

the following lemma implies, convex functions are directionally differentiable if we allow $\pm\infty$ as a limit.

Lemma 2.3 ([Bertsekas2003])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be convex. Then, the function $h_{x,d}\colon (0, +\infty) \to (-\infty, +\infty]$ defined by

$$h_{x,d}(\beta) \coloneqq \frac{h(x + \beta d) - h(x)}{\beta}$$

is non-decreasing. In particular, it follows that

$$h'(x; d) \leq h_{x,d}(\beta) \leq h(x + d) - h(x) \quad \textit{for all } x, d \in \mathbf{R}^p, \beta \in (0, 1].$$

On the other hand, for a proper and convex function $h\colon \mathbf{R}^p \to (-\infty, +\infty]$, we call $\xi \in \mathbf{R}^p$ a *subgradient* of h at $x \in \mathbf{R}^p$ if

$$h(y) - h(x) \geq \langle \xi, y - x \rangle \quad \text{for all } y \in \mathbf{R}^p,$$

and we write $\partial h(x)$ the *subdifferential* of h at x, i.e., the set of all subgradients of h at x. When h is differentiable at x, $\partial h(x)$ amounts to a singular $\{\nabla h(x)\}$.

2.6 The proximal operator and Moreau envelope

We suppose that $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ is closed, proper, and convex. Then, we define the *Moreau envelope* or *Moreau-Yosida regularization* $\mathcal{M}_h\colon \mathbf{R}^p \to \mathbf{R}$ by

$$\mathcal{M}_h(x) \coloneqq \min_{y \in \mathbf{R}^p} \left[h(y) + \frac{1}{2}\|x - y\|_2^2 \right]. \tag{2.9}$$

The minimization problem in (2.9) has a unique solution because of the strong convexity of its objective function. We call this solution the *proximal operator* and write it as

$$\mathbf{prox}_h(x) = \operatorname*{argmin}_{y \in \mathbf{R}^p} \left[h(y) + \frac{1}{2}\|x - y\|_2^2 \right]. \tag{2.10}$$

The proximal operator is non-expansive, i.e., $\|\mathbf{prox}_h(x) - \mathbf{prox}_h(y)\|_2 \leq \|x - y\|_2$ for any $x, y \in \mathbf{R}^p$. This also means that $\mathbf{prox}_h$ is 1-Lipschitz continuous. Moreover, when h is the indicator function (1.7) of a non-empty, closed, and convex set $C \subseteq \mathbf{R}^p$, we have

$$\mathbf{prox}_{\delta_C}(x) = \mathbf{proj}_C(x), \tag{2.11}$$

where $\mathbf{proj}_C$ is the projection onto C defined by (1.5). Even if h is non-differentiable, its Moreau envelope $\mathcal{M}_h$ is differentiable.

Theorem 2.1 ([Beck2017])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, $\mathcal{M}_h$ is 1-smooth and

$$\nabla \mathcal{M}_h(x) = x - \mathbf{prox}_h(x).$$

We also refer to the so-called second prox theorem and a corollary quickly derived from it.

Theorem 2.2 (Second prox theorem [Beck2017])
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, it follows that

$$\langle x - \mathbf{prox}_h(x), y - \mathbf{prox}_h(x)\rangle \leq h(y) - h(\mathbf{prox}_h(x)) \quad \textit{for all } x, y \in \mathbf{R}^p.$$

Corollary 2.1
Let $h\colon \mathbf{R}^p \to (-\infty, +\infty]$ be closed, proper, and convex. Then, we have

$$\|x - \mathbf{prox}_h(x)\|_2^2 \leq h(x) - h(\mathbf{prox}_h(x)) \quad \textit{for all } x \in \mathbf{R}^p.$$

2.7 Polyak-Łojasiewicz inequality and proximal-PL inequality

We focus on the following unconstrained optimization problem:

$$\min_{x \in \mathbf{R}^n} \quad f(x), \tag{2.12}$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable. Assume that (2.12) has an optimal solution, and let f^* denote the optimal function value. Then, we say that f satisfies the Polyak-Łojasiewicz (PL) inequality if there exists $\mu_f > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu_f (f(x) - f^*) \quad \text{for all } x \in \mathbf{R}^n. \tag{2.13}$$

Equation (2.13) is valid, for example, when f is strongly convex. Under (2.13), the steepest descent method [**Cauchy1847**] solving (2.12) converges linearly [**Polyak1963**].

On the other hand, we consider the composite optimization (1.6), supposing that F^* denotes the optimal function value. If there exists $\mu_{f,g} > 0$ such that

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu_{f,g}(F(x) - F^*) \quad \text{for all } x \in \mathbf{R}^n, \tag{2.14}$$

where

$$\mathcal{D}_g(x, \beta) := -2\beta \min_{y \in \mathbf{R}^n} \left[\langle \nabla f(x), y - x \rangle + g(y) - g(x) + \frac{\beta}{2}\|y - x\|_2^2 \right].$$

Like (2.13), the strong convexity of g is the sufficient condition of (2.14). With (2.14), the proximal gradient method described by Algorithm 1.1 for (1.6) converges linearly [**Karimi2016**].

2.8 Quasi-Féjer convergence

We define the concept of *quasi-Féjer convergence* and introduce a related theorem useful for the global convergence analysis.

Definition 2.1 (Quasi-Féjer convergence)
We say that $\{x^k\} \subseteq \mathbf{R}^p$ *is quasi-Féjer convergent to a non-empty set* $T \subseteq \mathbf{R}^p$ *if for all* $x \in T$ *there exists* $\{\varepsilon_k\} \subseteq \mathbf{R}_+$ *such that*

$$\|x^{k+1} - x\|_2^2 \leq \|x^k - x\|_2^2 + \varepsilon_k \quad \textit{and} \quad \sum_{\ell=0}^{\infty} \varepsilon_\ell < +\infty \quad \textit{for all } k = 0, 1, \ldots.$$

Theorem 2.3 ([Burachik1995])
If $\{x^k\}$ *is quasi-Féjer convergent to a non-empty set* $T \subseteq \mathbf{R}^p$*, then* $\{x^k\}$ *is bounded. Moreover, if an accumulation point* x^* *of* $\{x^k\}$ *belongs to* T*, then* $\lim_{k\to\infty} x^k = x^*$

2.9 Stability and sensitivity analysis

We consider the following parameterized optimization problem:

$$\min_{x \in X} \quad h(x, \xi), \tag{2.15}$$

depending on the parameter vector $\xi \in \Xi$. We assume that $X \subseteq \mathbf{R}^p$ and $\Xi \subseteq \mathbf{R}^q$ are non-empty and closed. Let us write the optimal value function of (2.15)

$$\phi(\xi) \coloneqq \inf_{x \in X} h(x, \xi) \tag{2.16}$$

and the associated set as

$$\Phi(\xi) \coloneqq \{x \in X \mid \phi(\xi) = h(x, \xi)\}.$$

The following proposition describes the directional differentiability of the optimal value function ϕ.

Proposition 2.1 ([Bonnans2000])
Let $\xi^0 \in \Xi$. Suppose that

(i) the function $h(x, \xi)$ is continuous on $X \times \Xi$;

(ii) there exist $\alpha \in \mathbf{R}$ and a compact set $C \subseteq X$ such that for every $\hat{\xi}$ near ξ^0, the level set $\mathbf{lev}_\alpha h(\cdot, \hat{\xi})$ is non-empty and contained in C;

(iii) for any $x \in X$, the function $h_x(\cdot) \coloneqq h(x, \cdot)$ is directionally differentiable at ξ^0;

(iv) if $\xi \in \Xi, t_k \searrow 0$, and $\{x^k\} \subseteq C$, then $\{x^k\}$ has an accumulation point $\bar{x}$ such that

$$\limsup_{k \to \infty} \frac{h(x^k, \xi^0 + t_k(\xi - \xi^0)) - h(x^k, \xi^0)}{t_k} \geq h'_{\bar{x}}(\xi^0; \xi - \xi^0).$$

Then, the optimal value function ϕ given by (2.16) is directionally differentiable at ξ^0 and

$$\phi'(\xi^0; \xi - \xi^0) = \inf_{x \in \Phi(\xi^0)} h'_x(\xi^0; \xi - \xi^0).$$

2.10 Pareto optimality

Let us introduce the concept of optimality for the multi-objective optimization problem (1.1).

Definition 2.2 (Pareto optimality and weak Pareto optimality)
For (1.1), we say that $x \in C$ is

(i) Pareto optimal *if there is no $y \in C$ such that $F(y) \leq F(x)$ and $F(y) \neq F(x)$;*

(ii) weakly Pareto optimal *if there does not exist* $y \in C$ *such that* $F(y) < F(x)$.

By definition, weak Pareto optimality contains Pareto optimality, though both definitions reduce to the usual optimality when $m = 1$. On the other hand, if the objective functions are non-convex, it is challenging to find Pareto minima or weak Pareto minima. In such cases, optimization algorithms aim to get Pareto stationary points defined as follows:

Definition 2.3 (Pareto stationarity)
Assume that F_i *is directionally differentiable for every* $i = 1, \ldots, m$, *and* C *is nonempty, closed, and convex. Then, we call* $x \in C$ Pareto stationary *if*

$$\max_{i=1,\ldots,m} F_i'(x; y - x) \geq 0 \quad \textit{for all } y \in C.$$

We state below the relation among the three concepts given by Definitions 2.2 and 2.3

Lemma 2.4
Suppose that F_i *is directionally differentiable for every* $i = 1, \ldots, m$, *and* C *is nonempty, closed, and convex. Then, the following three claims hold.*

(i) If $x \in C$ *is weakly Pareto optimal for* (1.1), *then* x *is Pareto stationary for* (1.1).

(ii) Let every $F_i, i = 1, \ldots, m$ *be convex. Then, all Pareto stationary points of* (1.1) *are weakly Pareto optimal for* (1.1).

(iii) Suppose that F_i *is strictly convex for any* $i = 1, \ldots, m$. *Then, every Pareto stationary point of* (1.1) *is Pareto optimal for* (1.1).

Proof. We prove each claim's contraposition.

Claim (i): Assume that $x \in C$ is not Pareto stationary. Then, Definition 2.3 shows that for some $y \in C$ we have $\max_{i=1,\ldots,m} F_i'(x; y - x) < 0$. By the definition (2.8) of the directional derivative, for a sufficiently small scalar $\beta > 0$, we obtain

$$\max_{i=1,\ldots,m} [F_i(x + \beta(y - x)) - F_i(x)] < 0,$$

which means that x is not weakly Pareto optimal from Definition 2.2 (ii).

Claim (ii): Suppose that $x \in C$ is not weakly Pareto optimal. Then, Definition 2.2 (ii) implies that there exists $y \in C$ such that $F_i(y) < F_i(x)$ for all $i = 1, \ldots, m$. Therefore, the convexity of F_i and Lemma 2.3 give

$$F'(x; y - x) \leq F_i(y) - F_i(x) < 0 \quad \text{for all } i = 1, \ldots, m.$$

Hence, we get

$$\max_{i=1,\ldots,m} F'(x; y - x) < 0,$$

which implies that x is not Pareto stationary from Definition 2.3.

Claim (iii): Suppose that $x \in C$ is not Pareto optimal. From Definition 2.2 (i), there exists $y \in C$ such that $F(y) \leq F(x)$ and $F(y) \neq F(x)$. Since F_i is strictly convex for every $i = 1, \ldots, m$, we have

$$F(x + \beta(y - x)) < F(x) + \beta(F(y) - F(x)) \quad \text{for all } \beta \in (0, 1).$$

Reducing $F(x)$ and dividing by β from both sides lead to

$$\frac{F(x + \beta(y - x)) - F(x)}{\beta} < F(y) - F(x) \leq 0.$$

Applying Lemma 2.3 to each component yields

$$F_i'(x; y - x) < F_i(y) - F_i(x) \leq 0,$$

which shows that x is not Pareto stationary. □

Chapter 3

Merit functions for multi-objective optimization

3.1 Introduction

This chapter considers the convex-constrained multi-objective optimization problems, i.e., (1.1) with C being non-empty, closed, and convex. It presents new merit functions for them, and discusses their properties mentioned in Section 1.3.

In detail, it proposes the following three merit functions for (1.1):

(i) the gap function for continuous multi-objective optimization;

(ii) the regularized gap function for convex multi-objective optimization;

(iii) the regularized and partially linearized gap function for composite multi-objective optimization.

In Table 3.1, we summarize the properties of those merit functions, which will be shown in the subsequent sections. There, 'Sol.' represents the types of Pareto solutions for (1.1) corresponding to the merit functions' minima (zero points). Moreover, 'SP,' 'LB,' and 'EB' indicate each F_i's sufficient conditions so that stationary points of the merit functions can solve (1.1), the merit functions are level-bounded, and the merit functions provide error bounds, respectively. The gap function (i) connects its minima and the weak Pareto solutions of (1.1) but does not have good properties in other aspects. The regularized gap function (ii) has better properties but requires the convexity of F_i. The regularized and partially linearized gap function (iii) relaxes the convexity assumption and is easy to compute for particular problems.

Table 3.1: Properties of our proposed merit functions

(a) Proposed merit functions and their properties

	Obj.	Sol.	Cont.	Diff.	SP	LB	EB
(i)	Cont.	WPO	LSC	×	×	LB	SgC
(ii)	Conv.		Cont.	DD	SC	Conv., LB	
(iii)	Comp.	PS			SC, C^2	Conv., LB, etc.	SgC, etc.

(b) Table of abbreviations

Obj.	Objective functions
Sol.	Solutions
Cont.	Continuity
Diff.	Differentiability
SP	Stationary points
LB	Level-boundedness
EB	Error bounds
Cont.	Continuity
Comp.	Composite
WPO	Weak Pareto optimality
PS	Pareto stationarity
LSC	Lower semicontinuity
DD	Directional differentiability
C^2	Twice continuously differentiable
SC	Strict convexity
SgC	Strong convexity

We summarize the structure of the rest of this chapter. Section 3.2 proposes different merit functions for multi-objective optimization with continuous, convex, and composite objectives, respectively, along with methods for evaluating the function values, the differentiability, and the stationary point properties. Furthermore, sufficient conditions for them to be level-bounded and to provide error bounds are given in Sections 3.4 and 3.5, respectively, introducing the extension of the proximal-PL inequality [**Chambolle2015**] to multi-objective problems.

3.2 Merit functions and their basic properties

This section proposes different types of merit functions for the multi-objective optimization (1.1), considering three cases: when the objective function F is continuous, it is convex, and it has a composite structure.

3.2.1 A gap function for continuous multi-objective optimization

First, we assume only continuity on F other than continuity and propose a gap function $u_\infty \colon C \to (-\infty, +\infty]$ as follows:

$$u_\infty(x) := \sup_{y \in C} \min_{i=1,\dots,m} [F_i(x) - F_i(y)]. \tag{3.1}$$

When F is linear, this merit function has already been discussed in [**Liu2009**], but here we consider the more general nonlinear cases. We now show that u_∞ is a merit function in the sense of weak Pareto optimality.

Theorem 3.1
Let u_∞ be defined by (3.1). Then, we have $u_\infty(x) \geq 0$ for all $x \in C$. Moreover, $x \in C$ is weakly Pareto optimal for (1.1) if and only if $u_\infty(x) = 0$.

Proof. Let $x \in C$. By the definition (3.1) of u_∞, we get

$$u_\infty(x) = \sup_{y \in C} \min_{i=1,\dots,m} [F_i(x) - F_i(y)] \geq \min_{i=1,\dots,m} [F_i(x) - F_i(x)] = 0.$$

On the other hand, again, considering the definition (3.1) of u_∞, we obtain

$$u_\infty(x) = 0 \iff \min_{i=1,\dots,m} [F_i(x) - F_i(y)] \leq 0 \quad \text{for all } y \in C.$$

This is equivalent to the existence of $i = 1, \ldots, m$ such that

$$F_i(x) - F_i(y) \leq 0 \quad \text{for all } y \in C,$$

i.e., there does not exist $y \in C$ such that

$$F_i(x) - F_i(y) > 0 \quad \text{for all } i = 1, \ldots, m,$$

which means that x is weakly Pareto optimal for (1.1) by Definition 2.2 (i). □

The following theorem is clear from the continuity of F_i.

Theorem 3.2

The function u_∞ defined by (3.1) is lower semi-continuous on C.

Theorems 3.1 and 3.2 imply that if $u_\infty(x^k) \to 0$ holds for some bounded sequence $\{x^k\}$, its accumulation points are weakly Pareto optimal. Thus, we can use u_∞ to measure the complexity of multi-objective optimization methods.

Moreover, Theorem 3.1 implies that we can get weakly Pareto optimal solutions via the following single-objective optimization problem:

$$\min_{x \in C} \quad u_\infty(x).$$

However, if F_i is not bounded from below on C, we cannot guarantee that u_∞ is finite-valued. Moreover, even if u_∞ is finite-valued, u_∞ does not preserve the differentiability of the original objective function F.

3.2.2 A regularized gap function for convex multi-objective optimization

Here, we suppose that each component F_i of the objective function F of (1.1) is convex. Then, we define a regularized gap function $u_\alpha \colon C \to \mathbf{R}$ with a given constant $\alpha > 0$, which overcomes the shortcomings mentioned at the end of the previous subsection, as follows:

$$u_\alpha(x) := \max_{y \in C} \min_{i=1,\ldots,m} \left[F_i(x) - F_i(y) - \frac{1}{2\alpha} \|x - y\|_2^2 \right]. \tag{3.2}$$

Note that the strong concavity of the function inside $\max_{y \in C}$ implies that u_α is finite-valued, and there exists a unique solution that attains this maximum in C.

Like u_∞, we can show that u_α is also a merit function in the sense of weak Pareto optimality.

Theorem 3.3
Let u_α be defined by (3.2) for some $\alpha > 0$. Then, we have $u_\alpha(x) \geq 0$ for all $x \in C$. Moreover, $x \in C$ is weakly Pareto optimal for (1.1) if and only if $u_\alpha(x) = 0$.

Proof. Let $x \in C$. The definition (3.2) of u_α yields

$$\begin{aligned} u_\alpha(x) &= \max_{y \in C} \min_{i=1,\ldots,m} \left[F_i(x) - F_i(y) - \frac{1}{2\alpha}\|x - y\|_2^2 \right] \\ &\geq \min_{i=1,\ldots,m} \left[F_i(x) - F_i(x) - \frac{1}{2\alpha}\|x - y\|_2^2 \right] = 0, \end{aligned}$$

which proves the first statement.

We now show the second statement. First, assume that $u_\alpha(x) = 0$. Then, (3.2) again gives

$$\min_{i=1,\ldots,m} \left[F_i(x) - F_i(y) - \frac{1}{2\alpha}\|x - y\|_2^2 \right] \quad \text{for all } y \in C.$$

Let $z \in C$ and $\beta \in (0, 1)$. Since the convexity of C implies that $x + \beta(z - x) \in C$, substituting $y = x + \alpha(z - x)$ into the above inequality, we get

$$\min_{i=1,\ldots,m} \left[F_i(x) - F_i(x + \beta(z - x)) - \frac{1}{2\alpha}\|\beta(z - x)\|_2^2 \right] \leq 0.$$

The convexity of F_i leads to

$$\min_{i=1,\ldots,m} \left[\beta(F_i(x) - F_i(z)) - \frac{1}{2\alpha}\|\beta(z - x)\|_2^2 \right] \leq 0.$$

Dividing both sides by β and letting $\beta \searrow 0$, we have

$$\min_{i=1,\ldots,m} [F_i(x) - F_i(z)] \leq 0.$$

Since z can take an arbitrary point in C, it follows from (3.1) that $u_\infty(x) = 0$. Therefore, from Theorem 3.1, x is weakly Pareto optimal.

Now, suppose that x is weakly Pareto optimal. Then, it follows again from Theorem 3.1 that $u_\infty(x) = 0$. It is clear that $u_\alpha(x) \leq u_\infty(x)$ from the definitions (3.1) and (3.2) of u_∞ and u_α, so we get $u_\alpha(x) = 0$. □

Let us now write

$$U_\alpha(x) \coloneqq \operatorname*{argmax}_{y\in C} \min_{i=1,\dots,m} \left[F_i(x) - F_i(y) - \frac{1}{2\alpha}\|x-y\|_2^2 \right]. \tag{3.3}$$

The optimality condition of the maximization problem associated with (3.2) and (3.3) gives

$$\frac{1}{\alpha}[x - U_\alpha(x)] \in \operatorname*{conv}_{i\in\mathscr{I}_\alpha(x)} \partial F_i(U_\alpha(x)) + N_C(U_\alpha(x)) \quad \text{for all } x \in C,$$

where N_C denotes the normal cone to the convex set C and

$$\mathscr{I}_\alpha(x) = \operatorname*{argmin}_{i=1,\dots,m}[F_i(x) - F_i(U_\alpha(x))].$$

Thus, for all $x \in C$ there exists $e(x) \in \Delta^m$, where Δ^m is the unit m-simplex given by (2.1), such that $e_j(x) = 0$ for all $j \notin \mathscr{I}_\alpha(x)$ and

$$\frac{1}{\alpha}\langle x - U_\alpha(x), z - U_\alpha(x)\rangle \le \sum_{i=1}^m e_i(x)[F_i(z) - F_i(U_\alpha(x))] \quad \text{for all } z \in C. \tag{3.4}$$

Then, we can also show the continuity of u_α and U_α without any particular assumption.

Theorem 3.4

For all $\alpha > 0$, u_α and U_α defined by (3.2) and (3.3) are locally Lipschitz continuous and locally Hölder continuous with exponent $1/2$ on C, respectively.

Proof. For any bounded set $\Omega \subseteq C$, let $x^1, x^2 \in \Omega$. Adding the two inequalities obtained by substituting $(x,z) = (x^1, U_\alpha(x^2))$ and $(x,z) = (x^2, U_\alpha(x^1))$ into (3.4), we get

$$\begin{aligned}
&\frac{1}{\alpha}\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\rangle \\
&\quad\le \sum_{i=1}^m [e_i(x^2) - e_i(x^1)]\,[F_i(U_\alpha(x^1)) - F_i(U_\alpha(x^2))] \\
&\quad= \sum_{i=1}^m e_i(x^1)\,[F_i(x^1) - F_i(U_\alpha(x^1))] + \sum_{i=1}^m e_i(x^2)\,[F_i(x^2) - F_i(U_\alpha(x^2))] \\
&\qquad+ \sum_{i=1}^m e_i(x^1)\,[F_i(U_\alpha(x^2)) - F_i(x^1)] + \sum_{i=1}^m e_i(x^2)\,[F_i(U_\alpha(x^1)) - F_i(x^2)].
\end{aligned}$$

Since $e(x) \in \Delta^m$ and $e_j(x) \neq 0$ for all $j \in \mathscr{I}_\alpha(x)$, we have

$$\begin{aligned}
&\frac{1}{\alpha}\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\rangle \\
&\quad = \min_{i=1,\dots,m}\left[F_i(x^1) - F_i(U_\alpha(x^1))\right] + \min_{i=1,\dots,m}\left[F_i(x^2) - F_i(U_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^m e_i(x^1)\left[F_i(U_\alpha(x^2)) - F_i(x^1)\right] + \sum_{i=1}^m e_i(x^2)\left[F_i(U_\alpha(x^1)) - F_i(x^2)\right]
\end{aligned}$$

Again using the fact that $e(x) \in \Delta^m$, we get

$$\begin{aligned}
&\frac{1}{\alpha}\langle U_\alpha(x^1) - U_\alpha(x^2) - (x^1 - x^2), U_\alpha(x^1) - U_\alpha(x^2)\rangle \\
&\quad \leq \sum_{i=1}^m e_i(x^2)\left[F_i(x^1) - F_i(U_\alpha(x^1))\right] + \sum_{i=1}^m e_i(x^1)\left[F_i(x^2) - F_i(U_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^m e_i(x^1)\left[F_i(U_\alpha(x^2)) - F_i(x^1)\right] + \sum_{i=1}^m e_i(x^2)\left[F_i(U_\alpha(x^1)) - F_i(x^2)\right] \\
&\quad = \sum_{i=1}^m \left[e_i(x^2) - e_i(x^1)\right]\left[F_i(x^1) - F_i(x^2)\right] \leq 2\max_{i=1,\dots,m}\left|F_i(x^1) - F_i(x^2)\right|.
\end{aligned}$$

Multiplying by α and adding $(1/4)\|x^1 - x^2\|^2$ in both sides of the inequality, it follows that

$$\left\|U_\alpha(x^1) - U_\alpha(x^2) - \frac{1}{2}(x^1 - x^2)\right\|_2^2 \leq \frac{1}{4}\|x^1 - x^2\|^2 + 2\alpha\max_{i=1,\dots,m}\left|F_i(x^1) - F_i(x^2)\right|.$$

Taking the square root of both sides, we obtain

$$\left\|U_\alpha(x^1) - U_\alpha(x^2) - \frac{1}{2}(x^1 - x^2)\right\|_2 \leq \sqrt{\frac{1}{4}\|x^1 - x^2\|^2 + 2\alpha\max_{i=1,\dots,m}|F_i(x^1) - F_i(x^2)|}.$$

Then, it follows from the triangle inequality that

$$\left\|U_\alpha(x^1) - U_\alpha(x^2)\right\|_2 \leq \frac{1}{2}\|x^1 - x^2\|_2 + \sqrt{\frac{1}{4}\|x^1 - x^2\|_2^2 + 2\alpha\max_{i=1,\dots,m}|F_i(x^1) - F_i(x^2)|}.$$

Since Lemma 2.1 implies that F_i locally Lipschitz continuous, there exists $L_{F_i}(\Omega) > 0$ such that

$$\left|F_i(x^1) - F_i(x^2)\right| \leq L_{F_i}(\Omega)\|x^1 - x^2\|_2. \tag{3.5}$$

Hence, the above two inequalities show U_α's local Hölder continuity with exponent $1/2$.

On the other hand, the definition (3.2) of u_α gives

$$\begin{aligned} u_\alpha(x^1) &= \max_{y\in C} \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(y) - \frac{1}{2\alpha}\|x^1 - y\|_2^2 \right] \\ &\geq \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) \right] - \frac{1}{2\alpha}\|x^1 - U_\alpha(x^2)\|_2^2. \end{aligned}$$

Reducing $u_\alpha(x^2)$ from both sides yields

$$u_\alpha(x^1) - u_\alpha(x^2) \geq \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{1}{2\alpha}\|x^1 - U_\alpha(x^2)\|_2^2 \right] - u_\alpha(x^2).$$

(3.2) and (3.3) lead to

$$\begin{aligned} u_\alpha(x^1) - u_\alpha(x^2) \geq & \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{1}{2\alpha}\|x^1 - U_\alpha(x^2)\|_2^2 \right] \\ & - \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(U_\alpha(x^2)) - \frac{1}{2\alpha}\|x^2 - U_\alpha(x^2)\|_2^2 \right]. \end{aligned}$$

From (2.2), we obtain

$$u_\alpha(x^1) - u_\alpha(x^2) \geq \min_{i=1,\dots,m} \left[F_i(x^1) - F_i(x^2) - \frac{1}{2\alpha}\langle x^1 + x^2 - 2U_\alpha(x^2), x^1 - x^2 \rangle \right].$$

Cauchy-Schwarz inequality and (3.5) implies

$$u_\alpha(x^1) - u_\alpha(x^2) \geq -\left[\max_{i=1,\dots,m} L_{F_i}(\Omega) + \frac{1}{2\alpha}\|x^1 + x^2 - 2U_\alpha(x^2)\|_2 \right] \|x^1 - x^2\|_2.$$

Since $U_\alpha(x)$ is bounded for $x \in \Omega$ due to the continuity, and the above inequality holds even if we interchange x^1 and x^2, we can show the local Lipschitz continuity of u_α. □

On the other hand, using the unit m-simplex Δ^m defined by (2.1), u_α given by (3.2) can also be expressed as

$$u_\alpha(x) = \max_{y\in C} \min_{e\in\Delta^m} \sum_{i=1}^m e_i \left[F_i(x) - F_i(y) - \frac{1}{2\alpha}\|x - y\|_2^2 \right].$$

We can see that C is convex, Δ^m is compact and convex, and the function inside $\min_{e\in\Delta^m}$ is convex for e and concave for y. Therefore, Sion's minimax theorem [**Sion1958**] leads to

$$\begin{aligned} u_\alpha(x) &= \min_{e\in\Delta^m} \max_{y\in C} \sum_{i=1}^m e_i \left[F_i(x) - F_i(y) - \frac{1}{2\alpha}\|x-y\|_2^2 \right] \\ &= \min_{e\in\Delta^m} \left[\sum_{i=1}^m e_i F_i(x) - \alpha^{-1}\, \mathcal{M}_{\alpha \sum_{i=1}^m e_i F_i + \delta_C}(x) \right], \end{aligned} \tag{3.6}$$

where $\mathcal{M}$ and δ_C denote the Moreau envelope and the indicator function defined by (1.7) and (2.9), respectively. Thus, we can evaluate u_α through the following m-dimensional, simplex-constrained, and convex optimization problem:

$$\begin{aligned} \min_{e\in\mathbf{R}^m} \quad & \sum_{i=1}^m e_i F_i(x) - \alpha^{-1}\, \mathcal{M}_{\alpha \sum_{i=1}^m e_i F_i + \delta_C}(x) \\ \text{s.t.} \quad & e \geq 0 \quad \text{and} \quad \sum_{i=1}^m e_i = 1. \end{aligned} \tag{3.7}$$

As the following theorem shows, (3.7) is also differentiable.

Theorem 3.5

The objective function of (3.7) *is continuously differentiable at every* $e \in \mathbf{R}^m$ *and*

$$\nabla_e \left[\sum_{i=1}^m e_i F_i(x) - \alpha^{-1}\, \mathcal{M}_{\alpha \sum_{i=1}^m e_i F_i + \delta_C}(x) \right] = F(x) - F\left(\mathbf{prox}_{\alpha \sum_{i=1}^m e_i F_i + \delta_C}(x) \right),$$

where **prox** *denotes the proximal operator* (2.10).

Proof. Define

$$h(y,e) \coloneqq \sum_{i=1}^m e_i F_i(y) + \frac{1}{2\alpha}\|x-y\|_2^2.$$

Clearly, h is continuous. Moreover, $h_y(\cdot) \coloneqq h(y,\cdot)$ is continuously differentiable and

$$\nabla_e h_y(e) = F(y).$$

Furthermore, $\mathbf{prox}_{\alpha\sum_{i=1}^m e_i F_i + \delta_C}(x) = \operatorname{argmin}_{y\in C} h(y,e)$ is also continuous at every $e \in \mathbf{R}^m$ from [**Rockafellar1998**]. Therefore, all the assumptions of Proposition 2.1 are satisfied. Since $\mathbf{prox}_{\alpha\sum_{i=1}^m e_i F_i + \delta_C}(x)$ is unique, we obtain the desired

result. □

Therefore, when $\mathbf{prox}_{\alpha\sum_{i=1}^m e_iF_i+\delta_C}(x)$ is easy to compute, we can solve (3.7) using well-known convex optimization techniques such as the interior point method [**Bertsekas1999**]. If $n \gg m$, this is usually faster than solving the n-dimensional problem directly to compute (3.2).

Let us now write the optimal solution set of (3.7) by

$$\mathcal{E}(x) \coloneqq \operatorname*{argmin}_{e\in\Delta^m}\left[\sum_{i=1}^m e_iF_i(x) - \alpha^{-1}\,\mathcal{M}_{\alpha\sum\limits_{i=1}^m e_iF_i+\delta_C}(x)\right]. \tag{3.8}$$

Then, we can show the (directional) differentiability of u_α, as in the following theorem.

Theorem 3.6
Let $x \in C$. For all $\alpha > 0$, the regularized gap function u_α defined by (3.2) is directionally differentiable at x and

$$u'_\alpha(x; z-x) = \inf_{e\in\mathcal{E}(x)}\left[\sum_{i=1}^m e_iF'_i(x; z-x) - \alpha^{-1}\left\langle x - \mathbf{prox}_{\alpha\sum\limits_{i=1}^m e_iF_i+\delta_C}(x), z-x\right\rangle\right]$$

for all $z \in C$, where $\mathcal{E}(x)$ is given by (3.8), and **prox** *denotes the proximal operator (2.10). In particular, if $\mathcal{E}(x)$ is a singleton, i.e., $\mathcal{E}(x) = \{e(x)\}$, and F_i is continuously differentiable at x, then u_α is continuously differentiable at x, and we have*

$$\nabla u_\alpha(x) = \sum_{i=1}^m e_i(x)\nabla F_i(x) - \alpha^{-1}\left(x - \mathbf{prox}_{\alpha\sum\limits_{i=1}^m e_i(x)F_i+\delta_C}\right).$$

Proof. Let

$$h(x,e) \coloneqq \sum_{i=1}^m e_iF_i(x) - \alpha^{-1}\,\mathcal{M}_{\alpha\sum\limits_{i=1}^m e_iF_i+\delta_C}(x).$$

Since $\mathcal{M}_{\alpha\sum_{i=1}^m e_iF_i+\delta_C}(x)$ is continuous at every $(x,e) \in C\times\Delta^m$ from [**Rockafellar1998**], h is also continuous on $C\times\Delta^m$. Moreover, Theorem 2.1 implies that for all $x, z \in C$ the function $h_e(\cdot) \coloneqq h(\cdot, e)$ has a directional derivative:

$$h'_e(x; z-x) = \sum_{i=1}^m e_iF'_i(x; z-x) - \alpha^{-1}\left\langle \mathbf{prox}_{\alpha\sum\limits_{i=1}^m e_iF_i+\delta_C}(x), z-x\right\rangle.$$

Because $\mathbf{prox}_{\alpha\sum_{i=1}^m e_i F_i + \delta_C}(x)$ is continuous at every $(x, e) \in C \times \Delta^m$ (cf. [**Rockafellar1998**]), $h'_e(x; z - x)$ is also continuous at every $(x, z, e) \in C \times C \times \Delta^m$. The discussion above and the compactness of Δ^m show that all assumptions of Proposition 2.1 are satisfied, so we get the desired result. □

From Theorems 3.3 and 3.6, the weakly Pareto optimal solutions for (1.1) are the globally optimal solutions of the following (directionally) differentiable single-objective optimization problem:

$$\min_{x \in C} \quad u_\alpha(x). \tag{3.9}$$

Since u_α is generally non-convex, (3.9) may have local optimal solutions or stationary points that are not globally optimal. As the following example shows, such stationary points are not necessarily Pareto stationary for (1.1).

Example 3.1
Let $m = 1, \alpha = 1, C = \mathbf{R}$ and $F_1(x) = |x|$. Then, we have

$$\mathcal{M}_{F_1}(x) = \begin{cases} x^2/2, & \text{if } |x| < 1, \\ |x| - 1/2, & \text{otherwise.} \end{cases}$$

Hence, we can evaluate u_1 as follows:

$$u_1(x) = \begin{cases} |x| - x^2/2, & \text{if } |x| < 1, \\ 1/2, & \text{otherwise.} \end{cases}$$

It is stationary for (3.9) at $|x| \geq 1$ and $x = 0$ but minimal only at $x = 0$. Furthermore, the stationary point of F_1 is only $x = 0$.

However, if we assume the strict convexity of each F_i, the stationary point of (3.9) is Pareto optimal for (1.1) and hence global optimal for (3.9). Note that this assumption does not assert the convexity of u_α.

Theorem 3.7
Suppose that F_i is strictly convex for all $i = 1, \ldots, m$. If $x \in C$ is a stationary point of (3.9), i.e.,

$$u'_\alpha(x; z - x) \geq 0 \quad \text{for all } z \in C,$$

then x is Pareto optimal for (1.1).

Proof. Let $e \in \mathcal{E}(x)$, where $\mathcal{E}(x)$ is given by (3.8). Then, Theorem 3.6 gives

$$\sum_{i=1}^{m} e_i F_i'(x; z-x) - \alpha^{-1} \left\langle x - \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x), z - x \right\rangle \geq 0 \quad \text{for all } z \in C.$$

Substituting $z = \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x)$ into the above inequality, we get

$$\sum_{i=1}^{m} e_i F_i' \left(x; \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) - x \right) + \alpha^{-1} \left\| x - \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) \right\|_2^2 \geq 0.$$

On the other hand, Theorem 2.2 yields

$$\left\| x - \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) \right\|_2^2 \leq \alpha \sum_{i=1}^{m} e_i \left[F_i(x) - F_i \left(\mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) \right) \right].$$

Combining the above two inequalities, we have

$$\sum_{i=1}^{m} e_i F_i' \left(x; \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) - x \right) \geq \sum_{i=1}^{m} e_i \left[F_i \left(\mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x) \right) - F_i(x) \right].$$

Since F_i is strictly convex for all $i = 1, \ldots, m$, the above inequality implies that $x = \mathbf{prox}_{\alpha \sum_{i=1}^{m} e_i F_i + \delta_C}(x)$, and hence $u_\alpha(x) = 0$. This means that x is Pareto optimal for (1.1) from the strict convexity of F_i, Lemmas 2.4 (i) and 2.4 (iii) and Theorem 3.3. □

3.2.3 A regularized and partially linearized gap function for composite multi-objective optimization

Now, let us consider the composite case (1.14). Since they are generally non-convex, we can regard them as a relaxation of the assumptions of the previous subsection. For (1.14), we propose a regularized and partially linearized gap function $w_\alpha \colon \mathbf{R}^n \to$

$\mathbf{R}$ with a given $\alpha > 0$ as follows:

$$w_\alpha(x) := \max_{y \in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha} \|x - y\|_2^2 \right]. \tag{3.10}$$

Like u_α, the convexity of g_i leads to the finiteness of w_α and the existence of a unique solution that attains $\max_{y \in C}$. As the following remark shows, w_α generalizes other kinds of merit functions.

Remark 3.1

(i) When $g_i = 0$, w_α corresponds to the regularized gap function (1.13) for vector variational inequality.

(ii) When $f_i = 0$, w_α matches u_α defined by (3.2).

As shown in the following theorem, w_α is a merit function in the sense of Pareto stationarity.

Theorem 3.8

Let w_α be given by (3.10) for some $\alpha > 0$. Then, we have $w_\alpha(x) \geq 0$ for all $x \in C$. Furthermore, $x \in C$ is Pareto stationary for (1.14) if and only if $w_\alpha(x) = 0$.

Proof. We first show the nonnegativity of w_α for all $\alpha > 0$. Let $x \in C$. The definition of w_α gives

$$\begin{aligned} w_\alpha(x) &= \sup_{y \in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha} \|x - y\|_2^2 \right] \\ &\geq \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - x \rangle + g_i(x) - g_i(x) - \frac{1}{2\alpha} \|x - x\|_2^2 \right] = 0. \end{aligned}$$

Let us prove the second statement. Assume that $w_\alpha(x) = 0$. Then, again using the definition of w_α, we get

$$\min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha} \|x - y\|_2^2 \right] \leq 0 \quad \text{for all } y \in C.$$

Let $z \in C$ and $\beta \in (0, 1)$. Since $C \subseteq \mathbf{R}^n$ is convex, $x, z \in C$ implies $x + \beta(z - x) \in C$. Therefore, by substituting $y = x + \beta(z - x)$ into the above inequality, we obtain

$$\min_{i=1,\dots,m} \left[-\langle \nabla f_i(x), \beta(z - x) \rangle + g_i(x) - g_i(x + \beta(z - x)) - \frac{1}{2\alpha} \|\beta(z - x)\|^2 \right] \leq 0.$$

Dividing both sides by β yields

$$\min_{i=1,\dots,m}\left[-\langle\nabla f_i(x), z-x\rangle - \frac{g_i(x+\beta(z-x))-g_i(x)}{\beta} - \frac{\alpha\beta}{2}\|z-x\|^2\right] \leq 0.$$

By taking $\beta \searrow 0$ and multiplying both sides by -1, we get

$$\max_{i=1,\dots,m} F_i'(x; z-x) \geq 0,$$

which means that x is Pareto stationary for (1.14).

Now, we prove the converse by contrapositive. Suppose that $w_\alpha(x) > 0$. Then, from the definition of w_α, there exists some $y \in C$ such that

$$\min_{i=1,\dots,m}\left[\langle\nabla f_i(x), x-y\rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha}\|x-y\|_2^2\right] > 0.$$

Since g_i is convex, we obtain

$$\min_{i=1,\dots,m}\left[\langle\nabla f_i(x), x-y\rangle - g_i'(x; y-x) - \frac{1}{2\alpha}\|x-y\|_2^2\right] > 0.$$

Thus, we have

$$\max_{i=1,\dots,m} F_i'(x; y-x) \leq -\frac{1}{2\alpha}\|x-y\|_2^2 < 0,$$

which shows that x is not Pareto stationary for (1.14). □

While u_∞ and u_α given by (3.1) and (3.2) are merit functions in the sense of weak Pareto optimality, w_α defined by (3.10) is a merit function only in the sense of Pareto stationarity. As indicated by the following example, even if $w_\alpha(x) = 0$, x is not necessarily weakly Pareto optimal for (1.14).

Example 3.2

Consider the single-objective function $F\colon \mathbf{R} \to \mathbf{R}$ *defined by* $F(x) \coloneqq f(x) + g(x)$, *where*

$$f(x) \coloneqq -x^2 \quad \textit{and} \quad g(x) \coloneqq 0,$$

and set $C = \mathbf{R}$. *Then, we have*

$$w_\alpha(0) = \max_{y\in\mathbf{R}}\left[f'(0)(0-y) + g(0) - g(y) - \frac{1}{2\alpha}(y-0)^2\right] = \max_{y\in\mathbf{R}}\left[-\frac{1}{2\alpha}y^2\right] = 0,$$

but $x = 0$ *is not global minimal (i.e., weakly Pareto optimal) for* F.

We now define the optimal solution mapping $W_\alpha\colon \mathbf{R}^n \to \mathbf{R}^n$ associated with (3.10) by

$$W_\alpha(x) \coloneqq \operatorname*{argmax}_{y\in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x-y\rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha}\|x-y\|_2^2\right]. \quad (3.11)$$

From the optimality condition of the maximization problem associated with (3.10) and (3.11), we obtain

$$\frac{1}{\alpha}[x - W_\alpha(x)] \in \operatorname*{conv}_{i\in\mathcal{I}_\alpha(x)} [\nabla f_i(x) + \partial g_i(W_\alpha(x))] + N_C(W_\alpha(x)) \quad \text{for all } x \in C,$$

where N_C is the normal cone to the convex set C and

$$\mathcal{I}_\alpha(x) \coloneqq \operatorname*{argmin}_{i=1,\dots,m} [\langle \nabla f_i(x), x - W_\alpha(x)\rangle + g_i(x) - g_i(W_\alpha(x))]. \quad (3.12)$$

Therefore, for any $x \in C$ there exists $\lambda(x)$ belonging to the unit m-simplex Δ^m defined by (2.1) such that $\lambda_j(x) = 0$ for all $j \notin \mathcal{I}_\alpha(x)$ and

$$\frac{1}{\alpha}\langle x - W_\alpha(x), z - W_\alpha(x)\rangle \le \sum_{i=1}^m \lambda_i(x)[\langle \nabla f_i(x), z - W_\alpha(x)\rangle + g_i(z) - g_i(W_\alpha(x))] \quad (3.13)$$

for all $z \in C$. Particularly, if we substitute $z = x$, we get

$$\frac{1}{\alpha}\|x - W_\alpha(x)\|_2^2 \le w_\alpha(x) + \frac{1}{2\alpha}\|x - W_\alpha(x)\|_2^2,$$

which reduces to

$$w_\alpha(x) \ge \frac{1}{2\alpha}\|x - W_\alpha(x)\|_2^2. \quad (3.14)$$

We can also show the continuity of w_α and W_α.

Theorem 3.9
For all $\alpha > 0$, w_α and W_α defined by (3.10) and (3.11) are continuous on C. Moreover, if every $\nabla f_i, i = 1, \dots, m$ is locally Lipschitz continuous, w_α and W_α are locally Lipschitz continuous and locally Hölder continuous with exponent $1/2$ on C, respectively.

Proof. Let Ω be a bounded subset of C and let $x^1, x^2 \in \Omega$. Adding the two inequalities gotten by substituting $(x, z) = (x^1, W_\alpha(x^2))$ and $(x, z) = (x^2, W_\alpha(x^1))$

into (3.13), we obtain

$$\begin{aligned}
&\frac{1}{\alpha}\langle W_\alpha(x^1) - W_\alpha(x^2) - (x^1 - x^2), W_\alpha(x^1) - W_\alpha(x^2)\rangle \\
&\quad \le \sum_{i=1}^{m} \lambda_i(x^1)\left[\langle \nabla f_i(x^1), x^1 - W_\alpha(x^1)\rangle + g_i(x^1) - g_i(W_\alpha(x^1))\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^2)\left[\langle \nabla f_i(x^2), x^2 - W_\alpha(x^2)\rangle + g_i(x^2) - g_i(W_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^1)\left[\langle \nabla f_i(x^1), W_\alpha(x^2) - x^1\rangle + g_i(W_\alpha(x^2)) - g_i(x^1)\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^2)\left[\langle \nabla f_i(x^2), W_\alpha(x^1) - x^2\rangle + g_i(W_\alpha(x^1)) - g_i(x^2)\right].
\end{aligned}$$

Since $\lambda_j(x) = 0$ for $j \in \mathcal{I}_\alpha(x)$, we have

$$\begin{aligned}
&\frac{1}{\alpha}\langle W_\alpha(x^1) - W_\alpha(x^2) - (x^1 - x^2), W_\alpha(x^1) - W_\alpha(x^2)\rangle \\
&\quad \le \min_{i=1,\dots,m}\left[\langle \nabla f_i(x^1), x^1 - W_\alpha(x^1)\rangle + g_i(x^1) - g_i(W_\alpha(x^1))\right] \\
&\qquad + \min_{i=1,\dots,m}\left[\langle \nabla f_i(x^2), x^2 - W_\alpha(x^2)\rangle + g_i(x^2) - g_i(W_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^1)\left[\langle \nabla f_i(x^1), W_\alpha(x^2) - x^1\rangle + g_i(W_\alpha(x^2)) - g_i(x^1)\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^2)\left[\langle \nabla f_i(x^2), W_\alpha(x^1) - x^2\rangle + g_i(W_\alpha(x^1)) - g_i(x^2)\right] \\
&\quad \le \sum_{i=1}^{m} \lambda_i(x^2)\left[\langle \nabla f_i(x^1), x^1 - W_\alpha(x^1)\rangle + g_i(x^1) - g_i(W_\alpha(x^1))\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^1)\left[\langle \nabla f_i(x^2), x^2 - W_\alpha(x^2)\rangle + g_i(x^2) - g_i(W_\alpha(x^2))\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^1)\left[\langle \nabla f_i(x^1), W_\alpha(x^2) - x^1\rangle + g_i(W_\alpha(x^2)) - g_i(x^1)\right] \\
&\qquad + \sum_{i=1}^{m} \lambda_i(x^2)\left[\langle \nabla f_i(x^2), W_\alpha(x^1) - x^2\rangle + g_i(W_\alpha(x^1)) - g_i(x^2)\right].
\end{aligned}$$

Therefore, simple calculations give

$$\begin{aligned}\frac{1}{\alpha}\left\|W_\alpha(x^1)-W_\alpha(x^2)\right\|_2^2 \le \frac{1}{\alpha}\langle W_\alpha(x^1)-W_\alpha(x^2),x^1-x^2\rangle \\ +\sum_{i=1}^{m}\left[\lambda(x^2)-\lambda(x^1)\right]\left[g_i(x^1)-g_i(x^2)+\left\langle\nabla f_i(x^1),x^1-x^2\right\rangle\right. \\ \left.-\left\langle\nabla f_i(x^1)-\nabla f_i(x^2),x^2\right\rangle\right] \\ +\sum_{i=1}^{m}\lambda_i(x^1)\left\langle\nabla f_i(x^1)-\nabla f_i(x^2),W_\alpha(x^2)\right\rangle+\sum_{i=1}^{m}\lambda_i(x^2)\left\langle\nabla f_i(x^2)-\nabla f_i(x^1),W_\alpha(x^1)\right\rangle.\end{aligned} \tag{3.15}$$

When $x^1 \to x^2$, the right-hand side tends to zero, which means the continuity of W_α on C. Therefore, from the definition, we can also say that w_α is continuous on C immediately.

Assume that each $\nabla f_i, i = 1,\dots,m$ is locally Lipschitz continuous. Since g_i is also locally Lipschitz continuous from Lemma 2.1, we can prove the local Hölder continuity of W_α from (3.15). On the other hand, the definitions (3.10) and (3.11)

of w_α and W_α give

$$
\begin{aligned}
& w_\alpha(x^1) - w_\alpha(x^2) \\
= & \min_{i=1,\dots,m} \left[\langle \nabla f_i(x^1), x^1 - W_\alpha(x^1)\rangle + g_i(x^1) - g_i(W_\alpha(x^1))\right] - \frac{1}{2\alpha}\|x^1 - W_\alpha(x^1)\|_2^2 \\
& - \max_{y\in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x^2), x^2 - y\rangle + g_i(x^2) - g_i(y) - \frac{1}{2\alpha}\|x^2 - y\|_2^2\right] \\
\le & \min_{i=1,\dots,m} \left[\langle \nabla f_i(x^1), x^1 - W_\alpha(x^1)\rangle + g_i(x^1) - g_i(W_\alpha(x^1))\right] - \frac{1}{2\alpha}\|x^1 - W_\alpha(x^1)\|_2^2 \\
& - \min_{i=1,\dots,m} \left[\langle \nabla f_i(x^2), x^2 - W_\alpha(x^1)\rangle + g_i(x^2) - g_i(W_\alpha(x^1))\right] + \frac{1}{2\alpha}\|x^2 - W_\alpha(x^1)\|_2^2 \\
\le & \max_{i=1,\dots,m} \left[\langle \nabla f_i(x^1) - \nabla f_i(x^2), x^1 - W_\alpha(x^1)\rangle + \langle \nabla f_i(x^2), x^1 - x^2\rangle + g_i(x^1) - g_i(x^2)\right] \\
& - \frac{1}{2\alpha}\langle x^1 - x^2, x^1 + x^2 - 2W_\alpha(x^1)\rangle \\
\le & \|x^1 - W_\alpha(x^1)\|_2 \max_{i=1,\dots,m} \|\nabla f_i(x^1) - \nabla f_i(x^2)\|_2 \\
& + \max_{i=1,\dots,m} \|\nabla f_i(x^2)\|_2 \|x^1 - x^2\|_2 + \max_{i=1,\dots,m} |g_i(x^1) - g_i(x^2)| \\
& + \frac{1}{2\alpha}\|x^1 + x^2 - 2W_\alpha(x^1)\|_2 \|x^1 - x^2\|_2,
\end{aligned}
$$

where the first inequality comes from (2.2), and the third inequality follows from the Cauchy-Schwarz inequality. The above inequality holds even if we interchange x^1 and x^2. Furthermore, $W_\alpha(x)$ and $\nabla f_i(x)$ are bounded for any $x \in \Omega$ due to their continuity. Therefore, local Lipschitz continuity of ∇f_i and g_i implies the local Lipschitz continuity of w_α. □

On the other hand, in the same way as the derivation of (3.6), Sion's minimax theorem [**Sion1958**] gives another representation of w_α for $\alpha > 0$ as follows:

$$
w_\alpha(x) = \min_{\lambda\in\Delta^m} \max_{y\in C} \sum_{i=1}^m \lambda_i \left[\langle \nabla f_i(x), x - y\rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha}\|x - y\|_2^2\right],
$$

where Δ^m denotes the unit m-simplex (2.1). Moreover, simple calculations show

that

$$\begin{aligned} w_\alpha(x) = \min_{\lambda\in\Delta^m} \Bigg\{ & \sum_{i=1}^m \lambda_i g_i(x) + \frac{1}{2\alpha}\left\|\sum_{i=1}^m \lambda_i \nabla f_i(x)\right\|_2^2 \\ & - \min_{y\in C}\left[\sum_{i=1}^m \lambda_i g_i(y) + \frac{1}{2\alpha}\left\|x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x) - y\right\|_2^2\right]\Bigg\} \\ = \min_{\lambda\in\Delta^m} \Bigg[& \sum_{i=1}^m \lambda_i g_i(x) + \frac{1}{2\alpha}\left\|\sum_{i=1}^m \lambda_i \nabla f_i(x)\right\|_2^2 \\ & - \alpha^{-1}\,\mathcal{M}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x)\right)\Bigg], \end{aligned}$$

where $\mathcal{M}$ and δ_C are given by (1.7) and (2.9), respectively. In other words, we can compute w_α via the following m-dimensional, simplex-constrained, and convex optimization problem:

$$\begin{aligned} \min_{\lambda\in\mathbf{R}^m} \quad & \sum_{i=1}^m \lambda_i g_i(x) + \frac{1}{2\alpha}\left\|\sum_{i=1}^m \lambda_i \nabla f_i(x)\right\|_2^2 \\ & - \alpha^{-1}\,\mathcal{M}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x)\right) \\ \text{s.t.} \quad & \lambda \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1. \end{aligned} \tag{3.16}$$

Moreover, the following theorem proves that (3.16) is differentiable.

Theorem 3.10

The objective function of (3.16) is continuously differentiable at every $\lambda \in \mathbf{R}^m$ and

$$\begin{aligned} \nabla_\lambda & \left[\sum_{i=1}^m \lambda_i g_i(x) + \frac{\alpha}{2}\left\|\sum_{i=1}^m \lambda_i \nabla f_i(x)\right\|_2^2 - \alpha^{-1}\,\mathcal{M}_{\alpha\sum_{i=1}^m g_i + \delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x)\right)\right] \\ & = g(x) - g\left(\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x)\right)\right) \\ & \quad - \mathcal{J}_f(x)\left(\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i \nabla f_i(x)\right) - x\right), \end{aligned}$$

where **prox** *is the proximal operator* (2.10)*, and* $\mathcal{J}_f(x)$ *is the Jacobian matrix at* x *given by* (2.7)*.*

Proof. Let

$$\theta(y,\lambda) \coloneqq \sum_{i=1}^{m} \lambda_i g_i(y) + \frac{\alpha}{2}\left\| x - \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) - y \right\|_2^2.$$

Then, θ is continuous, $\theta_y(\cdot) \coloneqq \theta(y,\cdot)$ is continuously differentiable, and

$$\nabla_\lambda \theta_y(\lambda) = g(y) + \mathcal{J}_f(x)\left(y - x + \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right).$$

Moreover, $\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}(x) = \operatorname{argmin}_{y\in C} \theta(y,\lambda)$ is also continuous at every $\lambda \in \mathbf{R}^m$ (cf. [**Rockafellar1998**]). The above discussion implies that every assumption in Proposition 2.1 is satisfied. Combined with the uniqueness of $\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}(x)$, we get

$$\begin{aligned}
&\nabla_\lambda \left[\alpha^{-1}\, \mathcal{M}_{\alpha \sum_{i=1}^{m} \lambda_i g_i + \delta_C}\left(x - \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right) \right] \\
&\qquad = g\left(\mathbf{prox}_{\alpha \sum_{i=1}^{m} \lambda_i g_i + \delta_C}\left(x - \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right) \right) \\
&\qquad + \mathcal{J}_f(x)\left(\mathbf{prox}_{\alpha \sum_{i=1}^{m} \lambda_i g_i + \delta_C}\left(x - \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right) - x + \alpha \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right).
\end{aligned}$$

On the other hand, we have

$$\nabla_\lambda \left[\sum_{i=1}^{m} \lambda_i g_i(x) + \frac{\alpha}{2}\left\| \sum_{i=1}^{m} \lambda_i \nabla f_i(x) \right\|_2^2 \right] = g(x) + \alpha \mathcal{J}_f(x) \sum_{i=1}^{m} \lambda_i \nabla f_i(x).$$

Adding the above two equalities, we obtain the desired result. □

Thus, like (3.7), (3.16) is solvable with convex optimization techniques such as the interior point method [**Bertsekas1999**] when we can quickly evaluate $\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}(\cdot)$. When $n \gg m$, this usually gives a faster way to compute w_α.

Example 3.3

(i) If $g_i(x) = 0$ *for all* $i = 1, \ldots, m$*, then* $\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i + \delta_C}$ *reduces to the projection onto* C *from* (2.11)*.*

(ii) *If* $g_i(x) = g_1(x)$ *for any* $i = 1, \dots, m$, *or if* $g_i(x) = g_1(x_{I_i})$ *and the index sets* $I_i \subseteq \{1, \dots, n\}$ *do not overlap each other, then* $\mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i}$ *is computable with each* $\mathbf{prox}_{g_i}$ *when* $C = \mathbf{R}^n$.

(iii) *Even if there is an overlap, we can compute* $\mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i}$ *for special functions. For example, when* $m = 2, g_1(x) = \|x\|_1, g_2(x) = \|x\|_2^2$, *and* $C = \mathbf{R}^n$, $\lambda_1 g_1(x) + \lambda_2 g_2(x)$ *is the elastic net [***Zou2005***]. It has a proximal operator in closed-form [***Parikh2014***].*

Now, define the optimal solution set of (3.16) by

$$\Lambda(x) = \operatorname*{argmin}_{\lambda \in \Delta^m} \left[\sum_{i=1}^m \lambda_i g_i(x) + \frac{\alpha}{2} \left\| \sum_{i=1}^m \lambda_i \nabla f_i(x) \right\|_2^2 - \alpha^{-1} \mathcal{M}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right]. \quad (3.17)$$

Then, in the same manner as Theorem 3.6, we obtain the following theorem.

Theorem 3.11

Let $x \in C$. *Assume that* f_i *is twice continuously differentiable at* x. *Then, for all* $\alpha > 0$, *the merit function* w_α *defined by* (3.10) *has a directional derivative*

$$w'_\alpha(x; z - x) = \inf_{\lambda \in \Lambda(x)} \left[\sum_{i=1}^m \lambda_i g_i'(x; z - x) - \alpha^{-1} \left\langle \left[I_n - \alpha \sum_{i=1}^m \lambda_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right] - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x), z - x \right\rangle \right]$$

for all $z \in C$, *where* **prox** *and* Λ *is given by* (2.10) *and* (3.17), *respectively, and* $I_n \in \mathbf{R}^{n \times n}$ *is the* n*-dimensional identity matrix. In particular, if* $\Lambda(x)$ *is a singleton, i.e.,* $\Lambda(x) = \{\lambda(x)\}$, *and* g_i *is continuously differentiable at* x, *then* w_α *is continuously*

differentiable at x, and we have

$$\nabla w_\alpha(x) = \sum_{i=1}^m \lambda_i(x)\nabla F_i(x)$$
$$-\alpha^{-1}\left[I_n - \alpha\sum_{i=1}^m \lambda_i(x)\nabla^2 f_i(x)\right]\left[x - \mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i(x)g_i+\delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i(x)\nabla f_i(x)\right)\right].$$

If the convex part g_i is the same regardless of i, we get the following corollary without assuming the differentiability of g_i.

Corollary 3.1

Let $x \in C$ and $\alpha > 0$. Assume that f_i is twice continuously differentiable at x and $g_i = g_1$ for all $i = 1, \ldots, m$, and recall that w_α and **prox** *be defined by (2.10) and (3.10), respectively. If $\Lambda(x)$ given by (3.17) is a singleton, i.e., $\Lambda(x) = \{\lambda(x)\}$, then the function $w_\alpha - g_1$ is continuously differentiable at x, and we have*

$$\nabla_x(w_\alpha(x) - g_1(x))$$
$$= -\alpha^{-1}\left[I_n - \alpha\sum_{i=1}^m \lambda_i(x)\nabla^2 f_i(x)\right]\left[x - \mathbf{prox}_{\alpha g_1+\delta_C}\left(x - \alpha\sum_{i=1}^m \lambda_i(x)\nabla f_i(x)\right)\right] + \sum_{i=1}^m \lambda_i(x)\nabla f_i(x).$$

Corollary 3.1 implies that, under certain conditions, the merit function $w_\alpha = (w_\alpha - g_1) + g_1$ is composite, i.e., the sum of a continuously differentiable function and a convex one.

Theorems 3.8 and 3.11 show that the Pareto stationary points for (1.14) are global optimal for the following directionally differentiable single-objective optimization problem:

$$\min_{x\in C} \quad w_\alpha(x). \tag{3.18}$$

Moreover, when the assumptions of Corollary 3.1 hold, we can apply first-order methods such as the proximal gradient method [**Fukushima1981**] to (3.18). On the other hand, if we consider Example 3.1 with $f_i = 0$, we can see that the stationary point for (3.18) is not necessarily Pareto stationary for (1.14). However, if f_i is convex and twice continuously differentiable, and F_i is strictly convex, then we can prove that every stationary point of (3.18) is Pareto optimal for (1.14), i.e., global

optimal for (3.9). Note that this assumption does not assert the convexity of w_α.

Theorem 3.12

Let $x \in C$ and $\alpha > 0$. Suppose that f_i is convex and twice continuously differentiable at x, and F_i is strictly convex for any $i = 1, \ldots, m$. If x is stationary for (3.18), i.e.,

$$w'_\alpha(x; z - x) \geq 0 \quad \textit{for all } z \in C,$$

then x is Pareto optimal for (1.14).

Proof. Let $z \in C$ and $\lambda \in \Lambda(x)$, where $\Lambda(x)$ is defined by (3.17). Then, it follows from Theorem 3.11 that

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i g_i'(x; z - x) \\
&\quad - \alpha^{-1} \left\langle \left[I_n - \alpha \sum_{i=1}^m \lambda_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right] \right. \\
&\qquad \left. - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x), z - x \right\rangle \geq 0.
\end{aligned}$$

Substituting $z = \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i F_i + \delta_C}(x)$, we have

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i F_i' \left(x; \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) - x \right) \\
&\quad + \alpha^{-1} \left\langle \left[I_n - \alpha \sum_{i=1}^m \lambda_i \nabla^2 f_i(x) \right] \left[x - \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right], \right. \\
&\qquad \left. x - \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right\rangle \geq 0.
\end{aligned}$$

Since the convexity of f_i implies that $\nabla^2 f_i(x)$ is positive semidefinite, we get

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i F_i' \left(x; \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) - x \right) \\
&\qquad + \alpha^{-1} \left\| \mathbf{prox}_{\alpha \sum_{i=1}^m \lambda_i g_i + \delta_C} \left(x - \alpha \sum_{i=1}^m \lambda_i \nabla f_i(x) \right) \right\|_2^2 \geq 0.
\end{aligned}$$

Therefore, with similar arguments used in the proof of Theorem 3.7, we obtain $x =$

$\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i g_i+\delta_C}(x-\alpha\sum_{i=1}^m \lambda_i \nabla f_i(x))$, and thus $w_\alpha(x)=0$. Since F_i is strictly convex, x is Pareto optimal for (1.14) from Lemma 2.4 (iii) and Theorem 3.8. □

3.3 Relation between different merit functions

This section assumes that the problem has a composite structure (1.14) and discusses the connection between the merit functions proposed in Section 3.2. First, we show some inequalities between different types of merit functions.

Theorem 3.13
Let u_∞, u_α, and w_α be defined by (3.1), (3.2) and (3.10), respectively, for all $\alpha>0$. Then, the following statements hold.

(i) If f_i is μ_{f_i}-convex for some $\mu_{f_i}\in\mathbf{R}$ and $\mu_f=\min_{i=1,\dots,m}\mu_{f_i}$, then we have

$$\begin{cases} u_\infty(x)\le w_{\mu^{-1}}(x) \quad \text{and} \quad u_{\alpha^{-1}}(x)\le w_{(\mu_f+\alpha)^{-1}}(x), & \text{if } \mu_f\ge 0,\\ u_{(-\mu_f+\alpha)^{-1}}(x)\le w_{\alpha^{-1}}(x), & \text{otherwise} \end{cases}$$

for all $\alpha>0$ and $x\in C$.

(ii) If ∇f_i is L_{f_i}-Lipschitz continuous for some $L_{f_i}>0$ and $L_f:=\max_{i=1,\dots,m}L_{f_i}$, then we get

$$u_{(L_f+\alpha)^{-1}}(x)\le w_{\alpha^{-1}}(x), \quad u_\infty(x)\ge w_{L_f^{-1}}(x), \quad \text{and} \quad u_{\alpha^{-1}}(x)\ge w_{(L_f+\alpha)^{-1}}(x)$$

for all $\alpha>0$ and $x\in C$.

Proof. Claim (i): Let $i=1,\dots,m$. The μ_{f_i}-convexity of f_i gives

$$f_i(x)-f_i(y)\le\langle\nabla f_i(x),x-y\rangle-\frac{\mu_{f_i}}{2}\|x-y\|_2^2.$$

By the definition of μ_f, we get

$$f_i(x)-f_i(y)\le\langle\nabla f_i(x),x-y\rangle-\frac{\mu_f}{2}\|x-y\|_2^2.$$

Thus, recalling (1.14), we have

$$F_i(x) - F_i(y) \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\mu_f}{2}\|x - y\|^2,$$
$$F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\mu_f + \alpha}{2}\|x - y\|^2,$$
$$F_i(x) - F_i(y) - \frac{-\mu_f + \alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x - y\|^2,$$

so the desired inequalities are clear from (3.1), (3.2) and (3.10).

Claim (ii): Let $i = 1, \ldots, m$. Suppose that ∇f_i is L_{f_i}-Lipschitz continuous. Then, Lemma 2.2 yields

$$|f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x \rangle| \leq \frac{L_{f_i}}{2}\|x - y\|_2^2.$$

By the definition of L_f, we have

$$|f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x \rangle| \leq \frac{L_f}{2}\|x - y\|_2^2.$$

This gives

$$F_i(x) - F_i(y) - \frac{L_f + \alpha}{2}\|x - y\|_2^2 \leq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{\alpha}{2}\|x - y\|_2^2,$$
$$F_i(x) - F_i(y) \geq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{L_f}{2}\|x - y\|_2^2,$$
$$F_i(x) - F_i(y) - \frac{\alpha}{2}\|x - y\|_2^2 \geq \langle \nabla f_i(x), x - y \rangle + g_i(x) - g_i(y) - \frac{L_f + \alpha}{2}\|x - y\|_2^2.$$

Therefore, we immediately get $u_{(L_f+\alpha)^{-1}}(x) \leq w_{\alpha^{-1}}(x)$, $u_\infty(x) \geq w_{L_f^{-1}}(x)$, and $u_{\alpha^{-1}}(x) \geq w_{(L_f+\alpha)^{-1}}(x)$ for all $x \in C$ by (3.1), (3.2) and (3.10). □

Second, we present the relation between coefficients and the proposed merit functions' values.

Theorem 3.14

Recall that w_{α_1} is defined by (3.10) for all $\alpha_1 > 0$. Let α_2 be an arbitrary scalar such that $\alpha_1 \geq \alpha_2$. Then, we get

$$w_{\alpha_2}(x) \leq w_{\alpha_1}(x) \leq \frac{\alpha_1}{\alpha_2} w_{\alpha_2}(x) \quad \textit{for all } x \in C.$$

Proof. Let $x \in C$. Since $\alpha_1 \geq \alpha_2 > 0$, the definition (3.10) of w_{α_1} and w_{α_2} clearly gives the first inequality. Thus, we prove the second one. From the definition (3.10) of w_{α_1}, we have

$$\begin{aligned} w_{\alpha_1}(x) &= \sup_{y\in C} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x-y\rangle + g_i(x) - g_i(y) - \frac{1}{2\alpha_1}\|x-y\|_2^2 \right] \\ &= \frac{\alpha_1}{\alpha_2} \sup_{y\in C} \min_{i=1,\dots,m} \left[\left\langle \nabla f_i(x), \frac{\alpha_2}{\alpha_1}(x-y) \right\rangle + \frac{\alpha_2}{\alpha_1}(g_i(x) - g_i(y)) - \frac{1}{2\alpha_2}\left\| \frac{\alpha_2}{\alpha_1}(x-y)\right\|_2^2 \right] \\ &\leq \frac{\alpha_1}{\alpha_2} \sup_{y\in C} \min_{i=1,\dots,m} \left[\left\langle \nabla f_i(x), \frac{\alpha_2}{\alpha_1}(x-y) \right\rangle + g_i(x) - g_i\left(x - \frac{\alpha_2}{\alpha_1}(x-y)\right) \right. \\ &\qquad \left. - \frac{1}{2\alpha_2}\left\| \frac{\alpha_2}{\alpha_1}(x-y)\right\|_2^2 \right] \end{aligned}$$

where the first inequality follows from the convexity of g_i. Since C is convex, $x, y \in C$ implies $x - (\alpha_2/\alpha_1)(x-y) \in C$. Therefore, from the definition (3.10) of w_{α_2}, we get

$$w_{\alpha_1}(x) \leq \frac{\alpha_1}{\alpha_2} w_{\alpha_2}(x).$$

□

Considering Remark 3.1 (ii), we get the following corollary.

Corollary 3.2

Assume that each component F_i of the objective function F of (1.1) is convex. Recall that u_{α_1} is defined by (3.2) for all $\alpha_1 > 0$. Let α_2 be an arbitrary scalar such that $\alpha_1 \geq \alpha_2$. Then, we get

$$u_{\alpha_2}(x) \leq u_{\alpha_1}(x) \leq \frac{\alpha_1}{\alpha_2} u_{\alpha_2}(x) \quad \textit{for all } x \in C.$$

3.4 Level-boundedness of the proposed merit functions

As we discussed in Section 2.2, we call a function *level-bounded* if every level set is bounded. This is an essential property because it ensures that the sequences generated by descent methods have accumulation points. We state below sufficient

conditions for the level-boundedness of the merit functions proposed in Section 3.2.

Theorem 3.15
Let u_∞, u_α, and w_α be defined by (3.1), (3.2) and (3.10), respectively, for all $\alpha > 0$. Then, the following claims hold.

(i) If F_i is level-bounded for all $i = 1, \ldots, m$, then u_∞ is level-bounded.

(ii) If F_i is convex and level-bounded for all $i = 1, \ldots, m$, then u_α is level-bounded for all $\alpha > 0$.

(iii) Suppose that F has the composite structure (1.14). If f_i is μ_{f_i}-convex for some $\mu_{f_i} \in \mathbf{R}$ or ∇f_i is L_{f_i}-Lipschitz continuous for some $L_{f_i} > 0$, and F_i is convex and level-bounded for all $i = 1, \ldots, m$, then w_α is level-bounded for all $\alpha > 0$.

Proof. Claim (i): Suppose, contrary to our claim, that u_∞ is not level-bounded. Then, there exists $c \in \mathbf{R}$ such that $\{x \in C \mid u_\infty(x) \leq c\}$ is unbounded. By the definition (3.2) of u_∞, the inequality $u_\infty(x) \leq c$ can be written as

$$\sup_{y \in C} \min_{i=1,\ldots,m} [F_i(x) - F_i(y)] \leq c.$$

This implies that for some fixed $z \in C$, there exists $j = 1, \ldots, m$ such that

$$F_j(x) \leq F_j(z) + c.$$

Therefore, it follows that

$$\{x \in C \mid u_\infty(x) \leq c\} \subseteq \bigcup_{j=1}^{m} \{x \in C \mid F_j(x) \leq F_j(z) + c\}.$$

Since F_i is level-bounded for all $i = 1, \ldots, m$, the right-hand side must be bounded, which contradicts the unboundedness of the left-hand side.

Claim (ii): Recall the definitions (2.1), (2.9) and (2.10) of $\Delta^m, \mathcal{M}$, and **prox**. Equation (3.6) gives

$$\begin{aligned} u_\alpha(x) &= \min_{\lambda\in\Delta^m}\left[\sum_{i=1}^m \lambda_i F_i(x) - \alpha^{-1}\,\mathcal{M}_{\alpha\sum_{i=1}^m \lambda_i F_i+\delta_C}(x)\right] \\ &= \min_{\lambda\in\Delta^m}\sum_{i=1}^m \lambda_i\left[F_i(x) - F_i\left(\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i F_i+\delta_C}(x)\right)\right. \\ &\qquad\qquad \left. -\frac{1}{2\alpha}\left\|x - \mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i F_i+\delta_C}(x)\right\|_2^2\right] \\ &\geq \frac{1}{2}\min_{\lambda\in\Delta^m}\sum_{i=1}^m \lambda_i\left[F_i(x) - F_i\left(\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i F_i+\delta_C}(x)\right)\right] \\ &= \frac{1}{2}\min_{i=1,\ldots,m}\left[F_i(x) - F_i\left(\mathbf{prox}_{\alpha\sum_{i=1}^m \lambda_i F_i+\delta_C}(x)\right)\right], \end{aligned}$$

where the inequality follows from Corollary 2.1. Therefore, with similar arguments given in the proof of claim (i), we can show the level-boundedness of u_α by contradiction.

Claim (iii): From Theorems 3.13 and 3.14, there exist some $\tau > 0$ and $\beta > 0$ such that $u_\beta(x) \leq \tau w_\alpha(x)$ for all $x \in C$. Since claim (ii) implies that u_β is level-bounded, w_α is also level-bounded. □

The following example indicates that our proposed merit functions are not necessarily level-bounded, even if F is level-bounded.

Example 3.4
Consider the bi-objective function $F\colon \mathbf{R} \to \mathbf{R}^2$ *with each component given by*

$$F_1(x) \coloneqq x^2, \quad F_2(x) \coloneqq 0.$$

Then, the gap function u_∞ *defined by* (3.2) *is written as*

$$\begin{aligned} u_\infty(x) &= \sup_{y\in\mathbf{R}} \min[F_1(x) - F_1(y), F_2(x) - F_2(y)] \\ &= \sup_{y\in\mathbf{R}} \min[(x^2 - y^2), 0] = 0. \end{aligned}$$

On the other hand, F *is level-bounded because* $\lim_{\|x\|_2\to\infty} F_1(x) = \infty$.

3.5 The multi-objective proximal PL inequality and error bounds

For the multi-objective composite problem (1.14), this section extends the proximal-PL inequality introduced in Section 2.7 and shows that it induces the proposed merit function's error bound. We first define the *multi-objective proximal-PL inequality*.

Definition 3.1 (Multi-objective proximal-PL inequality) *Assume that $f_i, i = 1, \dots, m$ is L_{f_i}-smooth with $L_{f_i} > 0$ and let $L_f := \max_{i=1,\dots,m} L_{f_i}$. We say that (1.14) satisfies the multi-objective proximal-PL inequality if there exists $\tau > 0$ such that*

$$w_{L_f^{-1}}(x) \geq \tau u_\infty(x) \quad \text{for all } x \in \mathbf{R}^n \tag{3.19}$$

with u_∞ and $w_{L_f^{-1}}$ given by (3.1) and (3.10).

If $m = 1$, (3.19) reduces to the proximal-PL inequality for scalar optimization (2.14). We state below some sufficient conditions for (3.19).

Proposition 3.1

(i) When f_i is μ_{f_i}-convex with $\mu_{f_i} > 0$, (3.19) holds with $\tau := \min(\mu_f/L_f, 1)$, where $\mu_f := \min_{i=1,\dots,m} \mu_{f_i}$.

(ii) Assume that $f_i(x) := h(A_i x)$ with some strongly convex function h_i and linear transformation A_i, and $g_i := \delta_{\mathcal{X}_i}$ with $\mathcal{X}_i$ being a polyhedral set and $\delta_{\mathcal{X}_i}$ given by (1.7). If each $\min_{x \in \mathbf{R}^n} F_i(x)$ has a nonempty set X_i^* for $i = 1, \dots, m$, then (3.19) holds with some constant τ.

Proof. Claim (i): Since f_i is strongly convex, Theorem 3.13 (i) gives

$$u_\infty(x) \leq w_\mu(x) \quad \text{for all } x \in \mathbf{R}^n.$$

Applying Theorem 3.14 to the above inequality implies

$$u_\infty(x) \leq \max\left(\frac{L_f}{\mu_f}, 1\right) w_{L_f^{-1}}(x) \quad \text{for all } x \in \mathbf{R}^n,$$

which means

$$w_{L_f^{-1}}(x) \geq \min\left(\frac{\mu_f}{L_f}, 1\right) u_\infty(x) \quad \text{for all } x \in \mathbf{R}^n.$$

Claim (ii): Since $\mathcal{X}_i$ is polyhedral, we can write it as $\{x \in \mathbf{R}^n \mid B_i x \leq c_i\}$ for some matrix B_i and vector c_i for $i = 1, \ldots, m$. We now show that for all $i = 1, \ldots, m$ there exists some z_i such that

$$X_i^* = \{x \in \mathbf{R}^n \mid B_i x \leq c_i \text{ and } A_i x = z_i\}.$$

To obtain a contradiction, suppose that there exists $x^1 \in X_i^*$ and $x^2 \in X_i^*$ such that $A_i x^1 \neq A_i x^2$. Clearly, we have $f_i(x^1) = f_i(x^2)$. Since h_i is strongly convex, we get

$$\begin{aligned} f_i(x^1) &= \frac{1}{2} f_i(x^1) + \frac{1}{2} f_i(x^2) = \frac{1}{2} h_i(A_i x^1) + \frac{1}{2} h_i(A_i x^2) \\ &> h_i\left(A_i\left(\frac{1}{2}x^1 + \frac{1}{2}x^2\right)\right) = f_i\left(\frac{1}{2}x^1 + \frac{1}{2}x^2\right), \end{aligned}$$

which contradicts the fact that $x^1 \in X_i^*$. Therefore, we can use Hoffman's error bound [**Hoffman1952**], and so there exists some $\rho_i > 0$ such that for any $x \in \mathbf{R}^n$, there exists $x_i^* \in X_i^*$ with

$$\|x - x_i^*\|_2 \leq \rho_i \left\| \max\left[\begin{pmatrix} B_i \\ A_i \\ -A_i \end{pmatrix} x - \begin{pmatrix} c_i \\ z_i \\ -z_i \end{pmatrix}, 0 \right] \right\|_2.$$

Note that we take the max operator componentwise on the right-hand side. Since $B_i x - c_i \leq 0$ for all $x \in \mathrm{dom}(F)$, we have

$$\|x - x_i^*\|_2 \leq \rho_i \left\| \max\left[\begin{pmatrix} A_i \\ -A_i \end{pmatrix} x - \begin{pmatrix} z_i \\ -z_i \end{pmatrix}, 0 \right] \right\|_2 \quad \text{for all } x \in \mathrm{dom}(F),$$

which yields

$$\|x - x_i^*\|_2^2 \leq \rho_i^2 \|A_i x - z_i\|_2^2 \quad \text{for all } x \in \mathrm{dom}(F).$$

Since $\mathbf{proj}_{X_i^*}(x) \in X_i^*$, it follows that

$$\left\| x - \mathbf{proj}_{X_i^*}(x) \right\|_2^2 \leq \|x - x_i^*\|_2^2 \leq \rho_i^2 \left\| A_i\left(x - \mathbf{proj}_{X_i^*}(x)\right) \right\|_2^2 \quad \text{for all } x \in \mathrm{dom}(F). \tag{3.20}$$

Now, suppose that $x \in \mathrm{dom}(F)$. From the definition (3.1) of u_∞, we get

$$\begin{aligned} u_\infty(x) &= \sup_{z \in \mathbf{R}^n} \min_{i=1,\dots,m} [F_i(x) - F_i(z)] \\ &\leq \min_{i=1,\dots,m} \sup_{z \in \mathbf{R}^n} [F_i(x) - F_i(z)] = \min_{i=1,\dots,m} \Big[F_i(x) - F_i\Big(\mathbf{proj}_{X_i^*}(x)\Big) \Big], \end{aligned}$$

where the second equality holds because $\mathbf{proj}_{X_i^*}(x) = \mathrm{argmin}_{x \in \mathbf{R}^n} F_i(x)$. Assuming that h_i is σ_{h_i}-convex with $\sigma_{h_i} > 0$, it follows that

$$\begin{aligned} u_\infty(x) &= \min_{i=1,\dots,m} \Bigg[\Big\langle \nabla h_i(A_i x), A_i\Big(x - \mathbf{proj}_{X_i^*}(x)\Big) \Big\rangle \\ &\qquad + g_i(x) - g_i\Big(\mathbf{proj}_{X_i^*}(x)\Big) - \frac{\sigma_{h_i}}{2} \Big\| A_i\Big(x - \mathbf{proj}_{X_i^*}(x)\Big) \Big\|_2^2 \Bigg] \\ &= \min_{i=1,\dots,m} \Bigg[\Big\langle \nabla f_i(x), x - \mathbf{proj}_{X_i^*}(x) \Big\rangle + g_i(x) - g_i\Big(\mathbf{proj}_{X_i^*}(x)\Big) \\ &\qquad - \frac{\sigma_{h_i}}{2} \Big\| A_i\Big(x - \mathbf{proj}_{X_i^*}(x)\Big) \Big\|_2^2 \Bigg]. \end{aligned}$$

Applying (3.20) to the above inequality leads to

$$\begin{aligned} u_\infty(x) \leq \min_{i=1,\dots,m} \Bigg[&\Big\langle \nabla f_i(x), x - \mathbf{proj}_{X_i^*}(x) \Big\rangle + g_i(x) - g_i\Big(\mathbf{proj}_{X_i^*}(x)\Big) \\ &- \frac{\sigma_{h_i}}{2\rho_i^2} \Big\| x - \mathbf{proj}_{X_i^*}(x) \Big\|_2^2 \Bigg]. \end{aligned}$$

Let $e \in \Delta^m$ with Δ^m given by (2.1). Since $\min_{i=1,\dots,m} v_i = \min_{e\in\Delta^m} \sum_{i=1}^m e_i v_i$ for any $v \in \mathbf{R}^m$, we get

$$\begin{aligned}
u_\infty(x) &\le \min_{e\in\Delta^m} \sum_{i=1}^m e_i \left[\left\langle \nabla f_i(x), x - \mathbf{proj}_{X_i^*}(x) \right\rangle + g_i(x) - g_i\Big(\mathbf{proj}_{X_i^*}(x)\Big) \right. \\
&\qquad \left. - \frac{\sigma_{h_i}}{2\rho_i^2} \left\| x - \mathbf{proj}_{X_i^*}(x) \right\|_2^2 \right] \\
&\le \min_{e\in\Delta^m} \sup_{z\in\mathbf{R}^n} \sum_{i=1}^m e_i \left[\langle \nabla f_i(x), x - z\rangle + g_i(x) - g_i(z) - \frac{\sigma_{h_i}}{2\rho_i^2} \|x - z\|_2^2 \right] \\
&= \sup_{z\in\mathbf{R}^n} \min_{e\in\Delta^m} \sum_{i=1}^m e_i \left[\langle \nabla f_i(x), x - z\rangle + g_i(x) - g_i(z) - \frac{\sigma_{h_i}}{2\rho_i^2} \|x - z\|_2^2 \right] \\
&= \sup_{z\in\mathbf{R}^n} \min_{i=1,\dots,m} \left[\langle \nabla f_i(x), x - z\rangle + g_i(x) - g_i(z) - \frac{\sigma_{h_i}}{2\rho_i^2} \|x - z\|_2^2 \right] \\
&\le w_{\rho_i^2/\min_{i=1,\dots,m} \sigma_{h_i}}(x),
\end{aligned}$$

where the first equality follows from the Sion's minimax theorem [**Sion1958**], and the third equality comes from the definition (3.10) of $w_{\rho_i^2/\min_{i=1,\dots,m} \sigma_{h_i}}$. Thus, Theorem 3.14 gives

$$u_\infty(x) \le \max\left(\frac{L_f \rho_i^2}{\min_{i=1,\dots,m} \sigma_{h_i}}, 1 \right) w_{L_f^{-1}}(x),$$

which completes the proof. □

We now show that the multi-objective proximal-PL inequality (3.19) leads to the error-bound property.

Theorem 3.16

Let $x \in \mathbf{R}^n$. Suppose that f_i is L_{f_i}-smooth with $L_{f_i} > 0$ for each $i = 1, \dots, m$, $L_f := \max_{i=1,\dots,m} L_{f_i}$, and the multi-objective proximal-PL inequality (3.19) holds with $\tau > 0$. Then, the trajectory $\left\{ W_{L_f^{-1}}^k(x) := \overbrace{W_{L_f^{-1}} \circ \cdots \circ W_{L_f^{-1}}}^{m}(x) \right\}$ converges linearly to a weakly Pareto optimal point x^* and

$$u_\infty(x) \ge \frac{\tau L_f}{8} \|x - x^*\|_2^2,$$

where u_∞ and $W_{L_f^{-1}}$ are given by (3.1) and (3.11), respectively.

Proof. Recall that u_∞ is non-negative due to Theorem 3.1. We have

$$\sqrt{u_\infty(x)} - \sqrt{u_\infty\Big(W_{L_f^{-1}}(x)\Big)} = \frac{u_\infty(x) - u_\infty\Big(W_{L_f^{-1}}(x)\Big)}{\sqrt{u_\infty(x)} + \sqrt{u_\infty\Big(W_{L_f^{-1}}(x)\Big)}}.$$

The definition (3.1) of u_∞ gives

$$u_\infty(x) - u_\infty\Big(W_{L_f^{-1}}(x)\Big) \geq \min_{i=1,\dots,m}\Big[F_i(x) - F_i\Big(W_{L_f^{-1}}(x)\Big)\Big] \geq w_{L_f^{-1}}(x),$$

where the second inequality follows from Lemma 2.2 and (3.10) and (3.11). Note that this inequality, together with (3.19), proves that $\left\{W^k_{L_f^{-1}}(x)\right\}$ converges linearly to zero. On the other hand, since $u_\infty(x) \geq u_\infty\Big(W_{L_f^{-1}}(x)\Big)$ because of Theorem 3.8 and the above inequality, we get

$$\sqrt{u_\infty(x)} + \sqrt{u_\infty\Big(W_{L_f^{-1}}(x)\Big)} \leq 2\sqrt{u_\infty(x)} \leq 2\sqrt{w_{L_f^{-1}}(x)/\tau},$$

where the second inequality comes from (3.19). Then, the above three inequalities show

$$\sqrt{u_\infty(x)} - \sqrt{u_\infty\Big(W_{L_f^{-1}}(x)\Big)} \geq \frac{w_{L_f^{-1}}(x)}{2\sqrt{w_{L_f^{-1}}(x)/\tau}} = \frac{1}{2}\sqrt{\tau w_{L_f^{-1}}(x)}.$$

Therefore, it follows from (3.14) that

$$\sqrt{u_\infty(x)} - \sqrt{u_\infty\Big(W_{L_f^{-1}}(x)\Big)} \geq \frac{\sqrt{\tau L_f}}{2\sqrt{2}}\Big\|x - W_{L_f^{-1}}(x)\Big\|_2.$$

More generally, we arrive at

$$\sqrt{u_\infty\Big(W^k_{L_f^{-1}}(x)\Big)} - \sqrt{u_\infty\Big(W^{k+1}_{L_f^{-1}}(x)\Big)} \geq \frac{\sqrt{\tau L_f}}{2\sqrt{2}}\Big\|W^k_{L_f^{-1}}(x) - W^{k+1}_{L_f^{-1}}(x)\Big\|_2$$

for all $k = 0, 1, \dots$. Adding up the above inequality from $k = k_1$ to $k = k_2 - 1$ yields

$$\sqrt{u_\infty\Big(W^{k_1}_{L_f^{-1}}(x)\Big)} - \sqrt{u_\infty\Big(W^{k_2}_{L_f^{-1}}(x)\Big)} \geq \frac{\sqrt{\tau L_f}}{2\sqrt{2}}\sum_{k=k_1}^{k_2-1}\Big\|W^k_{L_f^{-1}}(x) - W^{k+1}_{L_f^{-1}}(x)\Big\|_2.$$

Thus, the triangle inequality implies

$$\sqrt{u_\infty\left(W^{k_1}_{L_f^{-1}}(x)\right)} - \sqrt{u_\infty\left(W^{k_2}_{L_f^{-1}}(x)\right)} \geq \frac{\sqrt{\tau L_f}}{2\sqrt{2}}\left\|W^{k_1}_{L_f^{-1}}(x) - W^{k_2}_{L_f^{-1}}(x)\right\|_2. \tag{3.21}$$

As $k_1, k_2 \to \infty$, the left-hand side tends to zero. Therefore, the right-hand side also tends to zero because of the non-negativity of the norm. This means that $\{W^k_{L_f^{-1}}(x)\}$ is the Cauchy sequence, which is convergent to some weakly Pareto optimal point x^*. Substituting $k_1 = 0$ and $k_2 = \infty$ into (3.21) leads to

$$\sqrt{u_\infty(x)} \geq \frac{\sqrt{\tau L_f}}{2\sqrt{2}}\|x - x^*\|_2.$$

□

This theorem also presents the error-bound property of w_α and u_α for any $\alpha > 0$ because of (3.19) and Theorems 3.13 (ii) and 3.14.

3.6 Conclusions

In this chapter, we first proposed a gap function for (1.1) in the sense of weak Pareto optimality and showed its lower semicontinuity. We also defined a regularized gap function when F is convex and discussed its continuity, the way of evaluating it, its differentiability, and the properties of its stationary points. Furthermore, when each F_i is composite, we introduced a regularized and partially linearized gap function in the sense of Pareto stationarity and showed similar properties. In addition, we gave sufficient conditions for the proposed merit functions to be level-bounded and to provide error bounds, introducing the multi-objective proximal-PL inequality.

Chapter 4

A proximal gradient method for multi-objective optimization

4.1 Introduction

This chapter proposes the proximal gradient method for the unconstrained composite multi-objective optimization, i.e., (1.14) with $C = \mathbf{R}^n$. The proposed method generalizes Algorithm 1.1. Moreover, we analyze the proposed method's convergence rate using the merit functions (3.1) and (3.10) to measure the complexity.

We also observe that the problem and the proposed method have many applications. For example, when g_i is an indicator function of a non-empty, closed, and convex set S, (1.14) is equivalent to the optimization problems with constraints $x \in S$. Also, as seen in Section 4.5, we can deal with robust optimization problems. These problems include uncertain parameters and consist in optimizing under the worst scenario. Although the literature about robust optimization is vast, the studies about robust multi-objective optimization are relatively new [**Ehrgott2014**, **Fliege2014**, **Morishita2016**].

The outline of this chapter is as follows. Section 4.2 proposes the proximal gradient methods for unconstrained multi-objective optimization. We estimate the global convergence rates of the proposed method in Section 4.4. In Section 4.5, we apply the proposed method to robust optimization. Finally, we report some numerical experiments by solving robust multi-objective optimization problems in Section 4.6.

4.2 The algorithm

For given $x \in \operatorname{dom} F$ and $\alpha > 0$, we consider the following minimization problem:

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\alpha(z; x) \tag{4.1}$$

with

$$\varphi_\alpha(z; x) := \max_{i=1,\ldots,m} [\langle \nabla f_i(x), z - x \rangle + g_i(z) - g_i(x)] + \frac{1}{2\alpha} \|z - x\|_2^2.$$

The convexity of g_i implies that $\varphi_\alpha(\cdot; x)$ is strongly convex, so (4.1) always has a unique solution. Let us write such a solution as $p_\alpha(x)$ and let $\theta_\alpha(x)$ be its optimal function value, i.e.,

$$p_\alpha(x) := \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi_\alpha(z; x) \quad \text{and} \quad \theta_\alpha(x) := \min_{z \in \mathbf{R}^n} \varphi_\alpha(z; x). \tag{4.2}$$

Note that

$$p_\alpha = W_\alpha \quad \text{and} \quad \theta_\alpha = -w_\alpha \tag{4.3}$$

for w_α and W_α defined by (3.10) and (3.11). The following proposition shows that $p_\alpha(x)$ and $\theta_\alpha(x)$ help to characterize the Pareto stationarity of (1.14).

Lemma 4.1
Let p_α and θ_α be defined in (4.2). Then, the following claims hold.

(i) The following three conditions are equivalent: (a) $x \in \mathbf{R}^n$ is Pareto stationary for (1.14); (b) $p_\alpha(x) = x$; (c) $\theta_\alpha(x) = 0$.

(ii) The mappings p_α and θ_α are continuous. Moreover, if each $\nabla f_i, i = 1, \ldots, m$ is locally Lipschitz continuous, p_α and θ_α are locally Hölder continuous with exponent $1/2$ and locally Lipschitz continuous, respectively.

Proof. We can prove the claims immediately from Theorems 3.8 and 3.9. □

From Lemma 4.1, we can treat $\|p_\alpha(x) - x\|_\infty < \varepsilon$ for some $\varepsilon > 0$ as a stopping criterion. Let us present our proposed algorithm in Algorithm 4.1. We can consider the following three stepsize selection procedures. Note that every rule guarantees the objective functions' non-incrementality, i.e.,

$$F_i\left(x^{k+1}\right) \leq F_i\left(x^k\right) \quad \text{for all } i = 1, \ldots, m, k \geq 0. \tag{4.4}$$

Algorithm 4.1 The proximal gradient method for multi-objective optimization

Input: $x^0 \in \mathrm{int}(\mathrm{dom}(F)), \varepsilon > 0$

$k \leftarrow 0$
loop
  $z^k \leftarrow p_{\alpha_k}(x^k)$ with some stepsize $\alpha_k > 0$
  if $\|z^k - x^k\|_\infty < \varepsilon$ **then**
    return x^k
  end if
  $x^{k+1} \leftarrow x^k + s_k(z^k - x^k)$ with some stepsize $s_k \in (0, 1]$
  $k \leftarrow k + 1$
end loop

4.2.1 Armijo rule along the feasible direction

We fix $\alpha_k \coloneqq \alpha$ with some constant $\alpha > 0$ for every $k = 0, 1, \ldots$ and compute s_k by

$$s_k \coloneqq \xi^{j_k}, \tag{4.5}$$

where j_k is the smallest non-negative integer satisfying

$$F_i(x^k + \xi^{j_k}(z^k - x^k)) \leq F_i(x^k) + \rho\xi^{j_k}\left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right] \tag{4.6}$$

with predifined constants $\rho, \xi \in (0, 1)$ for each $i = 1, \ldots, m$. The following lemma demonstrates the existence of the stepsize s_k satisfying this rule.

Lemma 4.2
If $x \in \mathbf{R}^n$ is not Pareto stationary for (1.14), for all $\alpha > 0$ and $\rho \in (0, 1)$ there exists some $\bar{s} > 0$ such that

$$F_i(x + s(z - x)) < F_i(x) + \rho s\left[\theta_\alpha(x) - \frac{1}{2\alpha}\|z - x\|_2^2\right]$$

for all $i = 1, \ldots, m$ and $s \in (0, \bar{s}]$.

Proof. Let $s \in (0, 1]$ and $i = 1, \ldots, m$. Since g_i is convex, we have

$$g_i(x + s(p_\alpha(x) - x)) \leq g_i(x) + s[g_i(p_\alpha(x)) - g_i(x)].$$

Combined with (2.6), we get

$$\begin{aligned} &F_i(x + s(p_\alpha(x) - x)) \\ &\quad \leq F_i(x) + s\langle \nabla f_i(x), p_\alpha(x) - x\rangle + s[g_i(p_\alpha(x)) - g_i(x)] + o(s\|p_\alpha(x) - x\|_2) \end{aligned}$$

with $o\colon [0, +\infty) \to \mathbf{R}$ satisfying $\lim_{t\to\infty} o(t)/t = 0$. Then, (4.2) gives

$$F_i(x + s(p_\alpha(x) - x)) \leq F_i(x) + s\left[\theta_\alpha(x) - \frac{1}{2\alpha}\|p_\alpha(x) - x\|_2^2\right] + o(s\|p_\alpha(x) - x\|_2).$$

Since x is not Pareto stationary, Lemma 4.1 (i) gives $\theta_\alpha(x) < 0$. Thus, the fact that $\rho \in (0, 1)$ completes the proof. □

4.2.2 Sufficient decrease rule along the proximal arc

Suppose that ∇f_i is locally Lipschitz continuous for $i = 1, \ldots, m$. Let $s_k = 1$ for all $k = 0, 1, \ldots$ and $\alpha_{-1} > 0$. For each iteration, we find the smallest non-negative integer j_k such that

$$F_i(p_{\xi^{j_k}\alpha_{k-1}}(x^k)) \leq F_i(x^k) + \theta_{\xi^{j_k}\alpha_{k-1}}(x^k) \quad \text{for all } i = 1, \ldots, m \tag{4.7}$$

with some constant $\xi \in (0, 1)$ for any $i = 1, \ldots, m$ and set

$$\alpha_k = \xi^{j_k}\alpha_{k-1}. \tag{4.8}$$

4.2.3 Constant stepsize

When $\nabla f_i, i = 1, \ldots, m$ is L_{f_i}-Lipschitz continuous and $L_f := \max_{i=1,\ldots,m} L_{f_i}$, we set $s_k = 1$ and $\alpha_k = \alpha$ with $\alpha \in (0, 1/L_f]$ for each $i = 1, \ldots, m$. Then, Lemma 2.2 ensures that

$$F_i(p_\alpha(x^k)) \leq F_i(x^k) + \theta_\alpha(x^k) \quad \text{for all } i = 1, \ldots, m \text{ and } k = 0, 1, \ldots. \tag{4.9}$$

4.3 Convergence of the method

We first show the (classical) global convergence of the proposed algorithm.

Theorem 4.1
Every accumulation point of $\{x^k\}$ *generated by Algorithm 4.1 with any stepsize selection procedure given in Sections 4.2.1 to 4.2.3, if it exists, is Pareto stationary for (1.14).*

Proof. Assume that $\{x^{k_j}\}$ converges to $\bar{x}$. According to Lemma 4.1 (i), it suffices to check that $\theta_\alpha(\bar{x}) = 0$ for some $\alpha > 0$. Considering (4.4) and the existence of a subsequence of $\{F_i(x^k)\}$ converging to $F_i(\bar{x})$, we have

$$\lim_{k\to\infty} F_i(x^k) = F_i(\bar{x}). \tag{4.10}$$

Let us now prove the claim for each stepsize rule.

Armijo rule along the feasible direction: Equations (3.14), (4.3), (4.5) and (4.6) give

$$F_i(x^{k+1}) - F_i(x^k) \leq \rho s_k \left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right] \leq \rho s_k \theta_\alpha(x^k) \leq 0.$$

for any $i = 1, \ldots, m$. Combined with (4.10), we get

$$\lim_{k\to\infty} s_k \theta_\alpha(x^k) = 0. \tag{4.11}$$

If $\limsup_{k\to\infty} s_k > 0$, it is clear that $\theta_\alpha(\bar{x}) = 0$. We now suppose that $\lim_{k\to\infty} s_k = 0$. If we fix some positive integer q, we have $s_k < \xi^q$ for sufficiently large k. This means that the Armijo condition (4.6) does not hold for the stepsize ξ^q, i.e.,

$$F_{i_k}(x^k + \xi^q(z^k - x^k)) > F_{i_k}(x^k) + \rho\xi^q\left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right] \tag{4.12}$$

for some $i_k = 1, \ldots, m$. Since the number of values that i_k can take is finite, some subsequence of $\{i_k\}$ converges to some $\bar{i} = 1, \ldots, m$. Therefore, if we choose a proper subsequence and take the limit in (4.12), we obtain

$$F_{\bar{i}}(\bar{x} + \xi^q(p_\alpha(\bar{x}) - \bar{x})) \geq F_{\bar{i}}(\bar{x}) + \rho\xi^q\left[\theta_\alpha(\bar{x}) - \frac{1}{2\alpha}\|p_\alpha(\bar{x}) - \bar{x}\|_2^2\right].$$

Since q can take arbitrary positive integer values, Lemma 4.2 shows that $\bar{x}$ is Pareto stationary.

Sufficient decrease rule along the proximal arc: Similarly to the derivation of (4.11), the condition (4.7) leads to $\theta_{\alpha_k}(x^k) \to 0$ as $k \to \infty$. If $\alpha := \lim_{k\to\infty} \alpha_k > 0$,

it is easy to see that $\theta_\alpha(\bar{x}) = 0$. Assume that $\lim_{k\to\infty} \alpha_k = 0$. Similar to the last paragraph, fixing some positive integer q and considering that (4.7) does not hold, we get

$$F_{\bar{i}}(p_{\xi^q}(\bar{x})) \geq F_{\bar{i}}(\bar{x}) + \theta_{\xi^q}(\bar{x}).$$

Since q can take any positive integer, the locally Lipschitz continuity of ∇f_i shows that $\bar{x}$ is Pareto stationary.

Constant stepsize is clear from (4.9) and the previous paragraph. □

We now introduce the following assumption, standard in the analysis of descent methods for vector optimization [**Fliege2000**, **Grana-Drummond2004**, **Fukuda2011**].

Assumption 4.1
For all sequence $\{y^k\} \subseteq F(\mathbf{R}^n)$ *such that* $y^{k+1} \leq y^k, k = 0, 1, \ldots,$ *there exists* $x \in \mathbf{R}^n$ *satisfying* $F(x) \leq y^k, k = 0, 1, \ldots.$

Under convexity and reasonable assumptions, we can also prove the true convergence of iterates as follows:

Theorem 4.2
Suppose that f_i *is convex for* $i = 1, \ldots, m$ *and let* $\{x^k\}$ *generated by Algorithm 4.1 with the stepsize selection procedure given in Sections 4.2.1 to 4.2.3. Under Assumption 4.1,* $\{x^k\}$ *converges to some* $x^* \in T := \{x \in \mathbf{R}^n \mid F(x) \leq F(x^k), k = 0, 1, \ldots\}$ *and* x^* *is weakly Pareto optimal for* (1.14).

Proof. Let $x \in T$. We have

$$\begin{aligned} \|x^{k+1} - x\|_2^2 &= \|x^k - x\|_2^2 + \|x^{k+1} - x^k\|_2^2 + 2\langle x^k - x^{k+1}, x - x^k\rangle \\ &= \|x^k - x\|_2^2 + s_k^2\|z^k - x^k\|_2^2 - 2s_k\langle z^k - x^k, x - x^k\rangle, \\ &\leq \|x^k - x\|_2^2 + s_k\|z^k - x^k\|_2^2 - 2s_k\langle z^k - x^k, x - x^k\rangle \end{aligned} \tag{4.13}$$

where the second equality follows from line 7 of Algorithm 4.1, and the inequality comes from the fact that $s_k \in (0, 1]$. Recall that Δ^m and $\mathcal{I}_{\alpha_k}$ are defined by (2.1) and (3.12), respectively. Then, (3.13) implies that there exists $\lambda(x^k) \in \Delta^m$ such

that $\lambda_j(x^k) = 0$ for any $j \in \mathcal{I}_{\alpha_k}(x^k)$ and

$$\begin{aligned}\frac{1}{\alpha_k}\langle x^k - z^k, x - z^k\rangle &\leq \sum_{i=1}^{m} \lambda_i(x^k)\left[\langle \nabla f_i(x^k), x - z^k\rangle + g_i(x) - g_i(z^k)\right] \\ &= \sum_{i=1}^{m} \lambda_i(x^k)\left[\langle \nabla f_i(x^k), x - x^k\rangle + g_i(x) - g_i(x^k)\right] \\ &\quad + \sum_{i=1}^{m} \lambda_i(x^k)\left[\langle \nabla f_i(x^k), x^k - z^k\rangle + g_i(x^k) - g_i(z^k)\right].\end{aligned}$$

Since f_i is convex and $x \in T$, the first term in the right-hand side is non-negative. Therefore, (3.12) and (4.2) and line 3 of Algorithm 4.1 give

$$\frac{1}{\alpha_k}\langle x^k - z^k, x - z^k\rangle \leq -\theta_{\alpha_k}(x^k) + \frac{1}{2\alpha_k}\|z^k - x^k\|_2^2.$$

Combining the above inequality with (4.13) gives

$$\|x^{k+1} - x\|_2^2 \leq \|x^k - x\|_2^2 - 2\alpha_k s_k \theta_{\alpha_k}(x^k) = \|x^k - x\|_2^2 + 2\alpha_k s_k |\theta_{\alpha_k}(x^k)|.$$

Since $T \neq \emptyset$, it is easy to see that $\sum_{k=0}^{\infty} \alpha_k s_k |\theta_{\alpha_k}(x^k)| < +\infty$ holds for each stepsize selection rule. Hence, Definition 2.1 implies that $\{x^k\}$ is quasi-Féjer convergent to T, and thus $\{x^k\}$ is bounded due to Theorem 2.3. Let x^* be an accumulation point of $\{x^k\}$ and assume that $x^{k_j} \to x^*$. If we fix some non-negative integer $\bar{k}$, for sufficiently large j, we have

$$F(x^{k_j}) \leq F\left(x^{\bar{k}}\right).$$

Taking $j \to \infty$ yields that

$$F(x^*) \leq F\left(x^{\bar{k}}\right).$$

Since $\bar{k}$ can take any non-negative integer, x^* belongs to T. Use Theorem 2.3, and we can complete the proof. □

4.4 Convergence rate of the method

For the convergence rate analysis, we assume the Lipschitz gradient condition.

Assumption 4.2
Each f_i is L_{f_i}-smooth with $L_{f_i} > 0$ for $i = 1, \dots, m$. We write

$$L_f \coloneqq \max_{i=1,\dots,m} L_{f_i}. \tag{4.14}$$

4.4.1 The non-convex case

We present below the main theorem of this subsection.

Theorem 4.3
Assume that at least one of the functions $F_1, \dots, F_m$ is bounded from below. Then, under Assumption 4.2, Algorithm 4.1 with the stepsize selection rule given by any of Sections 4.2.1 to 4.2.3 generates a sequence $\{x^k\}$ such that $\{w_1(x^k)\}$ is summable, where w_1 is given by (3.10). In particular,

$$\liminf_{k\to\infty} (k \log k) w_1(x^k) = 0$$

and

$$\min_{0\le j\le k-1} w_1(x^j) = O(1/k)$$

with $O\colon [0, +\infty) \to \mathbf{R}$ such that $\limsup_{t\to\infty} O(t)/t < \infty$.

Before proving Theorem 4.3, we first show the existence of a uniform lower bound on the stepsize s_k when we adopt the Armijo rule along the feasible direction.

Lemma 4.3
In Algorithm 4.1 with the Armijo rule along the feasible direction under Assumption 4.2, the stepsize s_k satisfies

$$s_k \ge s_{\min} \coloneqq \min\left[\frac{2\xi(1-\rho)}{\alpha L_f}, 1\right] \quad \text{for all } k \ge 0.$$

Proof. Recall that $s_k = \xi^{j_k}$ for some non-negative integer j_k in (4.5). If $j_k = 0$, $s_k = 1$ clearly satisfies $s_k \ge s_{\min}$. Thus, suppose that $j_k \ge 1$. Since j_k is the smallest non-negative integer satisfying (4.6), there exists $i_k = 1, \dots, m$ such that

$$F_{i_k}(x^k + \xi^{j_k-1}(z^k - x^k)) > F_{i_k}(x^k) + \rho\xi^{j_k-1}\left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right].$$

On the other hand, $L_{f_{i_k}}$-Lipschitz continuity of ∇f_{i_k} and Lemma 2.2 give

$$\begin{aligned}
F_{i_k}(x^k + \xi^{j_k-1}(z^k - x^k)) &\leq F_{i_k}(x^k) + \langle \nabla f_{i_k}(x^k), \xi^{j_k-1}(z^k - x^k)\rangle \\
&\quad + g_{i_k}(x^k + \xi^{j_k-1}(z^k - x^k)) - g_{i_k}(x^k) + \frac{L_{f_{i_k}}}{2}\|\xi^{j_k-1}(z^k - x^k)\|_2^2 \\
&\leq F_{i_k}(x^k) + \xi^{j_k-1}[\langle \nabla f_{i_k}(x^k), z^k - x^k\rangle + g_{i_k}(z^k) - g_{i_k}(x^k)] + \frac{L_f}{2}\|\xi^{j_k-1}(z^k - x^k)\|_2^2 \\
&\leq F_{i_k}(x^k) + \xi^{j_k-1}\left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right] + \frac{L_f\xi^{2(j_k-1)}}{2}\|z^k - x^k\|_2^2,
\end{aligned}$$

where the second inequality comes from the convexity of g_{i_k} and (4.14), and the third one follows from (4.2) and line 7 of Algorithm 4.1. Combining the above two inequalities lead to

$$(1-\rho)\xi^{j_k-1}\left[\theta_\alpha(x^k) - \frac{1}{2\alpha}\|z^k - x^k\|_2^2\right] + \frac{L_f\xi^{2(j_k-1)}}{2}\|z^k - x^k\|_2^2 > 0.$$

Thus, (3.14) yields

$$-\frac{(1-\rho)\xi^{j_k-1}}{\alpha}\|z^k - x^k\|_2^2 + \frac{L_f\xi^{2(j_k-1)}}{2}\|z^k - x^k\|_2^2 > 0.$$

Therefore, we have

$$s_k = \xi^{j_k} > \frac{2\xi(1-\rho)}{\alpha L_f} \geq s_{\min}.$$

□

Similarly, when we employ the Sufficient decrease rule along the proximal arc, α_k has a uniform lower bound.

Lemma 4.4

In Algorithm 4.1 with the Sufficient decrease rule along the proximal arc under Assumption 4.2, the stepsize α_k satisfies

$$\alpha_k \geq \alpha_{\min} := \min\left[\frac{\xi}{L_f}, 1\right] \quad \textit{for all } k \geq 0$$

Proof. It is clear from Lemma 2.2. □

We now show the main theorem as follows:

Proof of Theorem 4.3. Armijo rule along the feasible direction: Let $k \geq 0$ and take i such that F_i is bounded from below. Equations (4.5) and (4.6) give

$$\begin{aligned} F_i\left(x^{k+1}\right) - F_i\left(x^k\right) &\leq \rho s_k \left[\theta_\alpha\left(x^k\right) - \frac{1}{2\alpha}\left\|z^k - x^k\right\|_2^2\right] \\ &\leq -\rho s_{\min} w_\alpha\left(x^k\right) \\ &\leq -\rho s_{\min} \max(1, \alpha^{-1}) w_1\left(x^k\right), \end{aligned}$$

where the second inequality follows from (4.3) and Lemma 4.3, and the third one comes from Theorem 3.14. Since F_i is bounded from below, adding up the inequality from $k = 0$ to $k = \infty$ lead to the summability of $\left\{w_1\left(x^k\right)\right\}$.

Sufficient decrease rule along the proximal arc: Let $k \geq 0$ and let $i = 1, \ldots, m$ be an index such that F_i is bounded from below. Lemma 4.4 imply that

$$F_i\left(x^{k+1}\right) - F_i\left(x^k\right) \leq -w_{\alpha_{\min}}\left(x^k\right).$$

The proof from here is the same as in the previous paragraph.

Constant stepsize: This case is likewise apparent from (4.9). □

Remark 4.1

*When $g_i = 0$ for all i, references [***Calderon2020***, **Fliege2019***, **Grapiglia2015***] present the convergence rate of various multi-objective optimization methods. However, they all evaluate the convergence rate with measures that depend on the subproblems or variables used in their algorithms. This means that the comparison in terms of complexity between different methods is not easy using those measures. However, Theorem 4.3 analyzes the convergence rate using the merit function w_1, which can be defined uniformly by (3.10) for multi-objective optimization problems with a structure like (1.14). In the subsequent discussions, our convergence rate analyses have similar advantages.*

4.4.2 The convex case

This subsection assumes the convexity of the objective functions of (1.14). More precisely, let us suppose the following.

Assumption 4.3

Let f_i and g_i be μ_{f_i}-convex and μ_{g_i}-convex, respectively, with $\mu_{f_i} \in \mathbf{R}$ and $\mu_{g_i} \geq 0$ for $i = 1, \ldots, m$. Write $\mu_f := \min_{i=1,\ldots,m} \mu_{f_i}$ and $\mu_g := \min_{i=1,\ldots,m} \mu_{g_i}$. Then, we

suppose that $\mu_f + \mu_g \geq 0$.

Then, we can show the following recursive relation, which is helpful for the subsequent discussion.

Lemma 4.5
Under Assumption 4.3, the following three statements hold:

(i) Algorithm 4.1 with the Armijo rule along the feasible direction generates $\{x^k\}$ *such that*

$$\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] + \frac{\rho(1+\alpha\mu_g)}{2\alpha}\left\|x^{k+1}-x\right\|_2^2$$
$$\leq (1-\rho s_k)\min_{i=1,\ldots,m}\left[F_i\left(x^{k}\right) - F_i(x)\right] + \frac{\rho[1+\alpha\mu_g-\alpha s_k(\mu_f+\mu_g)]}{2\alpha}\left\|x^{k}-x\right\|_2^2$$

for all $x \in \mathbf{R}^n$ *and* $k \geq 0$.

(ii) With the Sufficient decrease rule along the proximal arc, the sequence $\{x^k\}$ *generated by Algorithm 4.1 satisfies*

$$\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] + \frac{1+\alpha_k\mu_g}{2\alpha_k}\left\|x^{k+1}-x\right\|_2^2 \leq \frac{1-\alpha_k\mu_f}{2\alpha_k}\left\|x^{k}-x\right\|_2^2$$

for all $x \in \mathbf{R}^n$ *and* $k \geq 0$.

(iii) Let $\{x^k\}$ *be generated by Algorithm 4.1 with the Constant stepsize. Then, we have*

$$\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] + \frac{1+\alpha\mu_g}{2\alpha}\left\|x^{k+1}-x\right\|_2^2 \leq \frac{1-\alpha\mu_f}{2\alpha}\left\|x^{k}-x\right\|_2^2$$

for all $x \in \mathbf{R}^n$ *and* $k \geq 0$.

Proof. Claim (i): Let $x \in \mathbf{R}^n$. Equations (4.5) and (4.6) give

$$\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] \leq \min_{i=1,\ldots,m}\left[F_i\left(x^{k}\right) - F_i(x)\right] + \rho s_k\left[\theta_\alpha\left(x^k\right) - \frac{1}{2\alpha}\left\|z^k-x^k\right\|_2^2\right].$$

Let $\lambda\left(x^k\right) \in \Delta^m$ such that $\lambda_j\left(x^k\right) = 0$ for any $j \in \mathcal{I}_\alpha\left(x^k\right)$, where Δ^m and $\mathcal{I}_\alpha$ are defined by (2.1) and (3.12), respectively. Then, (4.2) and line 7 of Algorithm 4.1

yield

$$\min_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right)-F_i(x)\right]-\min_{i=1,\dots,m}\left[F_i\left(x^k\right)-F_i(x)\right] \leq \rho s_k\sum_{i=1}^m\lambda_i\left(x^k\right)\left[\left\langle\nabla f_i\left(x^k\right),z^k-x^k\right\rangle+g_i\left(z^k\right)-g_i\left(x^k\right)\right].$$

Since $\lambda\left(x^k\right)\in\Delta^m$, we have

$$\min_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right)-F_i(x)\right]-(1-\rho s_k)\min_{i=1,\dots,m}\left[F_i\left(x^k\right)-F_i(x)\right] \leq \rho s_k\sum_{i=1}^m\lambda_i\left(x^k\right)\left[\left\langle\nabla f_i\left(x^k\right),z^k-x^k\right\rangle+f_i\left(x^k\right)-f_i(x)+g_i\left(z^k\right)-g_i(x)\right].$$

It follows from the μ_{f_i}-convexity of f_i that

$$\begin{aligned}&\min_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right)-F_i(x)\right]-(1-\rho s_k)\min_{i=1,\dots,m}\left[F_i\left(x^k\right)-F_i(x)\right]\\ &\quad\leq \rho s_k\sum_{i=1}^m\lambda_i\left(x^k\right)\left[\left\langle\nabla f_i\left(x^k\right),z^k-x\right\rangle+g_i\left(z^k\right)-g_i(x)-\frac{\mu_{f_i}}{2}\left\|x^k-x\right\|_2^2\right]\\ &\quad\leq \rho s_k\sum_{i=1}^m\lambda_i\left(x^k\right)\left[\left\langle\nabla f_i\left(x^k\right),z^k-x\right\rangle+g_i\left(z^k\right)-g_i(x)\right]-\frac{\rho s_k\mu_f}{2}\left\|x^k-x\right\|_2^2,\end{aligned}$$

where the second inequality comes from the fact that $\lambda\left(x^k\right)\in\Delta^m$ and $\mu_f=\min_{i=1,\dots,m}\mu_{f_i}$. Without loss of generality, (3.13), g_i's μ_{g_i}-convexity, and the fact

that $\mu_g = \min_{i=1,\ldots,m} \mu_{g_i}$ lead to

$$\begin{aligned}
&\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] - (1-\rho s_k)\min_{i=1,\ldots,m}\left[F_i\left(x^{k}\right) - F_i(x)\right] \\
&\quad\leq -\frac{\rho s_k}{\alpha}\left\langle z^k - x^k, z^k - x\right\rangle - \frac{\rho s_k \mu_f}{2}\left\|x^k - x\right\|_2^2 - \frac{\rho s_k \mu_g}{2}\left\|z^k - x\right\|_2^2 \\
&\quad= -\frac{\rho}{\alpha}\left\langle x^{k+1} - x^k, x^k - x\right\rangle - \frac{\rho}{\alpha s_k}\left\|x^{k+1} - x^k\right\|_2^2 \\
&\qquad - \frac{\rho s_k \mu_f}{2}\|x^k - x\|_2^2 - \frac{\rho s_k \mu_g}{2}\left\|\frac{1}{s_k}(x^{k+1} - x^k) + x^k - x\right\|_2^2 \\
&\quad= -\frac{\rho(1+\alpha\mu_g)}{\alpha}\left\langle x^{k+1} - x^k, x^k - x\right\rangle - \frac{\rho(2s_k + \alpha\mu_g)}{2\alpha s_k}\left\|x^{k+1} - x^k\right\|_2^2 \\
&\qquad - \frac{\rho s_k(\mu_f + \mu_g)}{2}\left\|x^k - x\right\|_2^2 \\
&\quad= -\frac{\rho(1+\alpha\mu_g)}{2\alpha}\left(\left\|x^{k+1} - x\right\|_2^2 - \left\|x^k - x\right\|_2^2 - \left\|x^{k+1} - x^k\right\|_2^2\right) \\
&\qquad - \frac{\rho(2s_k + \alpha\mu_g)}{2\alpha s_k}\left\|x^{k+1} - x^k\right\|_2^2 - \frac{\rho s_k(\mu_f + \mu_g)}{2}\left\|x^k - x\right\|_2^2 \\
&\quad\leq -\frac{\rho(1+\alpha\mu_g)}{2\alpha}\left(\left\|x^{k+1} - x\right\|_2^2 - \left\|x^k - x\right\|_2^2\right) - \frac{\rho s_k(\mu_f + \mu_g)}{2}\left\|x^k - x\right\|_2^2,
\end{aligned}$$

where the first equality follows from line 7 of Algorithm 4.1, and the second inequality holds since $s_k \in (0, 1]$. The above inequality is equivalent to the desired one.

Claim (ii): Let $x \in \mathbf{R}^n$. Equations (4.7) and (4.8) yield

$$\min_{i=1,\ldots,m}\left[F_i\left(x^{k+1}\right) - F_i(x)\right] \leq \min_{i=1,\ldots,m}\left[F_i\left(x^{k}\right) - F_i(x)\right] + \theta_{\alpha_k}\left(x^k\right).$$

For the proof of claim (i), by replacing α with α_k, ρ with 1, s_k with 1, and z^k with x^{k+1}, and adding $(1/2\alpha_k)\|x^{k+1} - x^k\|_2^2$ to the right-hand side, we obtain the desired inequality.

Claim (iii): This claim is clear from claim (ii). □

Now, we show that $\left\{u_\infty\left(x^k\right)\right\}$ converges to zero with rate $O(1/k)$ with Algorithm 4.1 under the following assumption.

Assumption 4.4

There exists a bounded set $\Omega \subseteq \mathbf{R}^n$ such that for all $x \in \mathbf{lev}_{F(x^0)}(F)$ with $\mathbf{lev}_{F(x^0)}$ given by (2.5), some $z \in \Omega$ satisfies $F(z) \leq F(x)$.

Remark 4.2

(i) In single-objective cases, if the optimization problem has at least one optimal solution x^*, then $\Omega = \{x^*\}$ satisfies Assumption 4.4.

(ii) When the level set $\mathbf{lev}_{F(x^0)}(F)$ is bounded, Assumption 4.4 is also satisfied. For example, this is the case when F_i is strongly convex for at least one i.

Theorem 4.4

Under Assumptions 4.1 to 4.3, Algorithm 4.1 with the stepsize selection rule given in any of Sections 4.2.1 to 4.2.3 generates a sequence $\{x^k\}$ converging to x^* such that $\{\min_{i=1,\dots,m}[F_i(x^k) - F_i(x^*)]\}$ is summable, in particular

$$\liminf_{k\to\infty}(k\log k)\min_{i=1,\dots,m}\left[F_i\left(x^k\right) - F_i(x^*)\right] = 0$$

and

$$\limsup_{k\to\infty} k\min_{i=1,\dots,m}\left[F_i\left(x^k\right) - F_i(x^*)\right] < +\infty.$$

Supposing Assumption 4.4 additionally, we also have

$$u_\infty\left(x^k\right) = O(1/k) \quad \text{for all } k \geq 1$$

with $O\colon [0,+\infty) \to \mathbf{R}$ satisfying $\limsup_{t\to 0} O(t)/t < +\infty$ and u_∞ given by (3.1).

Proof. Let $T \coloneqq \left\{x \in \mathbf{R}^n \;\middle|\; F(x) \leq F\left(x^k\right), k = 0, 1, \dots\right\}$. From Theorem 4.2, $\{x^k\}$ converges to $x^* \in T$.

We first prove the claim for the Armijo rule along the feasible direction. Since $\mu_g \geq 0$ and $\mu_f + \mu_g \geq 0$, Lemmas 4.5 (i) and 4.3 yield

$$\begin{aligned}\min_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i(x^*)\right] + \frac{\rho}{2\alpha}\left\|x^{k+1} - x^*\right\|_2^2 \\ \leq (1 - \rho s_{\min})\min_{i=1,\dots,m}\left[F_i\left(x^k\right) - F_i(x^*)\right] + \frac{\rho}{2\alpha}\left\|x^k - x^*\right\|_2^2\end{aligned}$$

Adding up the above inequality from $k = 0$ to $k = \ell$, we obtain

$$\begin{aligned}
&\rho s_{\min} \sum_{k=0}^{\ell} \min_{i=1,\dots,m} \left[F_i\left(x^k\right) - F_i(x^*)\right] \\
&\leq \min_{i=1,\dots,m} \left[F_i(x^0) - F_i(x^*)\right] + \frac{\rho}{2\alpha}\left\|x^0 - x^*\right\|_2^2 \\
&\quad - \min_{i=1,\dots,m} \left[F_i\left(x^{\ell+1}\right) - F_i(x^*)\right] - \frac{\rho}{2\alpha}\left\|x^{\ell+1} - x^*\right\|_2^2 \\
&\leq \min_{i=1,\dots,m} \left[F_i(x^0) - F_i(x^*)\right] + \frac{\rho}{2\alpha}\left\|x^0 - x^*\right\|_2^2,
\end{aligned}$$

which means the summability of $\left\{\min_{i=1,\dots,m}\left[F_i\left(x^k\right) - F_i(x^*)\right]\right\}$. We now suppose Assumption 4.4. Since (4.4) implies $F\left(x^\ell\right) \leq F\left(x^k\right)$ for all $k = 0, \dots, \ell$, the above inequality holds even if we replace x^* by $z \in \Omega$ such that $F(z) \leq F\left(x^\ell\right)$. Again using the relation $F\left(x^\ell\right) \leq F\left(x^k\right)$, we have

$$\rho s_{\min}(\ell + 1) \min_{i=1,\dots,m} \left[F_i(x^\ell) - F_i(z)\right] \;\leq\; \min_{i=1,\dots,m} \left[F_i(x^0) - F_i(z)\right] + \frac{\rho}{2\alpha}\left\|x^0 - z\right\|_2^2.$$

Therefore, we get

$$\min_{i=1,\dots,m} \left[F_i(x^\ell) - F_i(z)\right] \leq \frac{\min_{i=1,\dots,m} \left[F_i(x^0) - F_i(z)\right] + \frac{\rho}{2\alpha}\left\|x^0 - z\right\|_2^2}{\rho s_{\min}(\ell + 1)}.$$

Since $x^\ell \in \mathbf{lev}_{F(x^0)}(F)$, Assumption 4.4 implies that

$$u_\infty(x^\ell) = \sup_{z \in \mathbf{R}^n} \min_{i=1,\dots,m} \left[F_i(x^\ell) - F_i(z)\right] = \sup_{z \in \Omega} \min_{i=1,\dots,m} \left[F_i(x^\ell) - F_i(z)\right]$$

Due to the boundedness of Ω, we conclude that $u_\infty(x^\ell) = O(1/\ell)$.

We can likewise show the claim for the Sufficient decrease rule along the proximal arc by Lemmas 4.5 (ii) and 4.4 and for the Constant stepsize by Lemma 4.5 (iii). □

4.4.3 The case that the multi-objective proximal-PL inequality holds

This subsection analyzes the convergence rate of Algorithm 4.1 under the multi-objective proximal-PL condition defined by Definition 3.1.

Theorem 4.5

Suppose that Assumption 4.2 and Equation (3.19) hold with a constant $\tau > 0$. Then, $\{x^k\}$ generated by Algorithm 4.1 with the stepsize selection rule given in any of Sections 4.2.1 to 4.2.3 converges linearly to a weakly Pareto optimal solution x^* of (1.1), i.e.,

$$\|x^k - x^*\|_2 = O(\exp(-rk))$$

for some $r > 0$ with $O\colon [0, +\infty) \to \mathbf{R}$ satisfying $\limsup_{t\to 0} O(t)/t < +\infty$.

Proof. Similarly to the proof of Theorem 4.3, we have

$$F_i\left(x^{k+1}\right) - F_i\left(x^k\right) \leq -\hat{c} w_1\left(x^k\right)$$

with some constant $\hat{c} > 0$. Hence, Theorem 3.14 implies that

$$F_i\left(x^{k+1}\right) - F_i\left(x^k\right) \leq -c w_{L_f^{-1}}\left(x^k\right)$$

with some $c > 0$. Thus, the multi-objective proximal PL inequality (3.19) shows

$$F_i\left(x^{k+1}\right) - F_i\left(x^k\right) \leq -c\tau u_\infty\left(x^k\right).$$

Adding $F_i\left(x^k\right) - F_i(x)$ to both sides gives

$$F_i\left(x^{k+1}\right) - F_i(x) \leq F_i\left(x^k\right) - F_i(x) - c\tau u_\infty\left(x^k\right).$$

Therefore, it follows that

$$\sup_{x\in\mathbf{R}^n} \min_{i=1,\ldots,m} \left[F_i\left(x^{k+1}\right) - F_i(x)\right] \leq \sup_{x\in\mathbf{R}^n} \min_{i=1,\ldots,m} \left[F_i\left(x^k\right) - F_i(x)\right] - c\tau u_\infty\left(x^k\right),$$

which is equivalent to

$$u_\infty\left(x^{k+1}\right) \leq (1 - c\tau) u_\infty\left(x^k\right).$$

Applying this inequality recursively, we get

$$u_\infty\left(x^k\right) \leq (1 - c\tau)^k u_\infty(x^0).$$

From Theorem 3.16, there exists a weakly Pareto optimal point x^* and

$$\frac{\tau L_f}{8}\left\|x^k - x^*\right\|_2^2 \leq (1 - c\tau)^k u_\infty(x^0),$$

which completes the proof. □

4.5 Application to robust multi-objective optimization

Now, let us apply the proposed algorithms to robust multi-objective optimization. Here, we suppose that the problems include uncertain parameters. Moreover, suppose that we can estimate the set of these uncertain parameters. Then, we try to optimize by considering the worst scenario. We observe that studies about robust multi-objective optimization is relatively new [**Ehrgott2014**, **Fliege2014**, **Morishita2016**].

Here, we consider the convex function g_i defined as follows:

$$g_i(x) \coloneqq \max_{u \in \mathcal{U}_i} \hat{g}_i(x, u). \tag{4.15}$$

We call $\mathcal{U}_i \subseteq \mathbf{R}^n$ an *uncertainty set*. From now on, we assume that $\mathcal{U}_i \subset \mathbf{R}^n$ is convex, and $\hat{g}_i \colon \mathbf{R}^n \times \mathbf{R}^n \to (-\infty, +\infty]$ is closed, proper, and convex for x. It is easy to see that g_i is also closed, proper, and convex. However, g_i is not necessarily differentiable even if $\hat{g}_i$ is differentiable. First, let us reformulate the subproblem (4.1) by using an extra variable $\gamma \in \mathbf{R}$ as

$$\begin{aligned} \min_{\gamma, z} \quad & \gamma + \frac{1}{2\alpha}\|z - x\|_2^2 \\ \text{s.t.} \quad & \langle \nabla f_i(x), z - x \rangle + g_i(z) - g_i(x) \leq \gamma, \quad i = 1, \dots, m. \end{aligned}$$

Note that g_i is not easy to calculate; thus, the subproblem is challenging to solve. When $\hat{g}_i$ and $\mathcal{U}_i$ have some particular structure, the constraints can be written as explicit formulae using the duality of (4.15). Now, assume that the dual problem of the maximization problem (4.15) is written as follows:

$$\begin{aligned} \min_{w^i} \quad & \tilde{g}_i(x, w^i) \\ \text{s.t.} \quad & w^i \in \tilde{\mathcal{U}}_i(x), \end{aligned}$$

where $\tilde{g}_i \colon \mathbf{R}^n \times \mathbf{R}^m \to (-\infty, +\infty]$ and $\tilde{\mathcal{U}}_i \colon \mathbf{R}^n \to 2^{\mathbf{R}^m}$. If the strong duality holds,

then we see that the subproblem (4.1) is equivalent to

$$\begin{aligned} \min_{\gamma,z,w^1,\ldots,w^m} \quad & \gamma + \frac{1}{2\alpha}\|z-x\|_2^2 \\ \text{s.t.} \quad & \langle \nabla f_i(x), z-x\rangle + \tilde{g}_i(z,w^i) - g_i(x) \le \gamma, \\ & w^i \in \tilde{\mathcal{U}}_i(z), \quad i=1,\ldots,m. \end{aligned} \tag{4.16}$$

When $\tilde{g}_i$ and $\tilde{\mathcal{U}}_i$ have some explicit form, this problem is tractable. As we mention below, we can convert the above subproblem to some well-known convex optimization problems in this case. This idea can also be seen in [**Ben-tal1998**]. In the following, we will introduce some robust multi-objective optimization problems where the subproblems can be written as quadratic programming, second-order cone programming, or semi-definite programming problems.

4.5.1 Linearly constrained quadratic programming

Suppose that $\hat{g}_i(x,u) = \langle x,u\rangle$ and $\mathcal{U}_i = \{u \in \mathbf{R}^n \mid A_i u \le b^i\}$, where $A_i \in \mathbf{R}^{d\times n}$ and $b^i \in \mathbf{R}^d$, that is, $\hat{g}_i$ is linear in x, and $\mathcal{U}_i$ is a polyhedron. Suppose also that $\mathcal{U}_i$ is nonempty and bounded. Then, we can rewrite (4.15) as the following linear programming problem:

$$\begin{aligned} \max_{u} \quad & \langle x,u\rangle \\ \text{s.t.} \quad & A_i u \le b^i. \end{aligned} \tag{4.17}$$

Its dual problem is given by

$$\begin{aligned} \min_{w} \quad & \langle b^i, w\rangle \\ \text{s.t.} \quad & A_i^\top w = x, \\ & w \ge 0. \end{aligned}$$

Since the strong duality holds, we can convert the subproblem (4.1) (or, equivalently (4.16)) to a linearly constrained quadratic programming problem:

$$\begin{aligned} \min_{\gamma,z,w^1,\ldots,w^m} \quad & \gamma + \frac{1}{2\alpha}\|z-x\|_2^2 \\ \text{s.t.} \quad & \langle \nabla f_i(x), z-x\rangle + \langle b^i, w^i\rangle - g_i(x) \le \gamma, \\ & A_i^\top w^i = z, \\ & w^i \ge 0, \quad i=1,\ldots,m. \end{aligned} \tag{4.18}$$

4.5.2 Second-order cone programming

Suppose that $\hat{g}_i(x,u) = \langle x, u\rangle$ and $\mathcal{U}_i = \{a^i + P_i v \in \mathbf{R}^n \mid \|v\|_2 \le 1, v \in \mathbf{R}^n\}$, where $a^i \in \mathbf{R}^n$ and $P_i \in \mathbf{R}^{n\times n}$, that is, $\hat{g}_i$ is once again linear for x and $\mathcal{U}_i$ is an ellipsoid. Then, for all $i = 1,\ldots,m$ we have

$$\begin{aligned} g_i(x) &= \max_{u\in\mathcal{U}_i} \hat{g}_i(x,u) \\ &= \max_{v:\|v\|_2\le 1} \left\langle a^i + P_i v, x\right\rangle \\ &= \left\langle a^i, x\right\rangle + \max_{v:\|v\|_2\le 1} \left\langle P_i^\top x, v\right\rangle. \end{aligned}$$

If $P_i^\top x = 0$, then $\max_{v:\|v\|_2\le 1} \left\langle P_i^\top x, v\right\rangle = 0 = \left\|P_i^\top x\right\|_2$. If $P_i^\top x \ne 0$, then $v = P_i^\top x / \|P_i^\top x\|_2$ is a solution of $\max_{v:\|v\|_2\le 1} \left\langle P_i^\top x, v\right\rangle$, and hence $\max_{v:\|v\|_2\le 1} \left\langle P_i^\top x, v\right\rangle = \left\|P_i^\top x\right\|_2$. Consequently, we have

$$g_i(x) = \left\langle a^i, x\right\rangle + \left\|P_i^\top x\right\|_2.$$

Therefore, by introducing slack variables $\gamma \in \mathbf{R}$ and $\tau \in \mathbf{R}$, the subproblem (4.1) can be written as

$$\begin{aligned} \min_{\tau,\gamma,z} \quad & \tau \\ \text{s.t.} \quad & \left\langle \nabla f_i(x) + a^i, z-x\right\rangle + \left\|P_i^\top z\right\|_2 - \left\|P_i^\top x\right\|_2 \le \gamma, \quad i=1,\ldots,m, \\ & \gamma + \frac{1}{2\alpha}\|z-x\|_2^2 \le \tau. \end{aligned}$$

Note that convex quadratic constraints can be converted to second-order cone constraints. Using the expression given in [**Alizadeh2003**], we get the following second-

order cone programming problem (SOCP):

$$\begin{aligned} \min_{\tau,\gamma,z} \quad & \tau \\ \text{s.t.} \quad & \begin{bmatrix} -\langle \nabla f_i(x) + a^i, z - x \rangle + \gamma + \left\| P_i^\top x \right\|_2 \\ P_i^\top z \end{bmatrix} \in \mathcal{K}_{n+1}, \\ & \begin{bmatrix} 1 - \gamma + \tau \\ 1 + \gamma - \tau \\ \sqrt{2/\alpha}(z - x) \end{bmatrix} \in \mathcal{K}_{n+2}, \end{aligned} \tag{4.19}$$

where $\mathcal{K}_q := \{(y_0, \bar{y}) \in \mathbf{R} \times \mathbf{R}^{q-1} \mid y_0 \geq \|\bar{y}\|_2\}$ is the second-order cone in $\mathbf{R}^q$. The above SOCP can be solved efficiently with an interior point method [**Alizadeh2003**].

4.5.3 Semi-definite programming

Suppose that $\hat{g}_i(x, u) = \langle x + u, A_i(x + u) \rangle$ and $\mathcal{U}_i = \{a^i + P_i v \in \mathbf{R}^n \mid \|v\|_2 \leq 1\}$, where $A_i \in \mathbf{R}^{n \times n}$ and $A_i \succeq O, a^i \in \mathbf{R}^n$ and $P_i \in \mathbf{R}^{n \times n}$. Then, there exists a matrix $M_i \in \mathbf{R}^{n \times n}$ such that $A_i = M_i M_i^\top$. Note that $\hat{g}_i$ is convex quadratic and $\mathcal{U}_i$ is an ellipsoid. Here, without loss of generality we can assume that A is a symmetric matrix since $\langle x + u, A_i(x + u) \rangle = \left\langle x + u, \tilde{A}_i(x + u) \right\rangle$, where $\tilde{A}_i := \left(A_i + A_i^\top\right)/2$. Then, $g_i(x)$ can be given as

$$g_i(x) = \max_{v : \|v\|_2 \leq 1} \left\langle x + a^i + P_i v, A_i\left(x + a^i + P_i v\right) \right\rangle. \tag{4.20}$$

Since problem (4.20) is a maximization problem of a convex function, it is not a convex optimization problem. Fortunately, it can be seen as a subproblem of a trust region method, so its optimal value $g_i(x)$ can be obtained efficiently. Considering (4.20), we observe that

$$g_i(z) = \max_{v : \|v\|_2 \leq 1} \left\langle z + a^i + P_i v, A_i\left(z + a^i + P_i v\right) \right\rangle. \tag{4.21}$$

From [**Beck2006**], the Lagrangian dual of the maximization problem (4.21) is given by

$$\begin{aligned} \min_{\beta,w} \quad & -w \\ \text{s.t.} \quad & \begin{bmatrix} -P_i^\top A_i P_i & -P_i^\top A_i\big(z+a^i\big) \\ -\big(z+a^i\big)^\top A_i^\top P_i & -\big\langle z+a^i, A_i\big(z+a^i\big)\big\rangle - w \end{bmatrix} \succeq \beta \begin{bmatrix} -I_n & 0 \\ 0 & 1 \end{bmatrix}, \\ & \beta \geq 0, \end{aligned} \tag{4.22}$$

where I_n stands for the identity matrix of dimension n. Let (β^*, w^*) be an optimal solution of (4.22) and assume that $\dim(\ker(A_i + \beta^* I_n)) \neq 1$. Since both (4.21) and (4.22) have strictly feasible solutions and $I_n \succ O$, then the strong duality holds from [**Beck2006**]. Therefore, recalling (4.16), the subproblem (4.1) is equivalent to

$$\begin{aligned} \min_{\gamma,z,w,\beta} \quad & \gamma + \frac{1}{2\alpha}\|z-x\|_2^2 \\ \text{s.t.} \quad & \langle \nabla f_i(x), z-x\rangle - w_i - g_i(x) \leq \gamma, \\ & \begin{bmatrix} -P_i^\top A_i P_i + \beta_i I_n & -P_i^\top A_i\big(z+a^i\big) \\ -\big(z+a^i\big)^\top A_i^\top P_i & -\big\langle z+a^i, A_i\big(z+a^i\big)\big\rangle - w_i - \beta_i \end{bmatrix} \succeq O, \\ & \beta_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Now, by using slack variables $\tau \in \mathbf{R}$ and $\zeta^i \in \mathbf{R}$ and converting the convex quadratic constraints to second-order cone ones, we get the following semi-definite program-

ming problem:

$$
\begin{aligned}
\min_{\tau,\beta,w,\gamma,z} \quad & \tau \\
\text{s.t.} \quad & \langle \nabla f_i(x), z - x \rangle - w_i - g_i(x) \leq \gamma, \\
& \begin{bmatrix} 1 - \gamma + \tau \\ 1 + \gamma - \tau \\ \sqrt{\dfrac{2}{\alpha}}(z - x) \end{bmatrix} \in \mathcal{K}_{n+2}, \\
& \begin{bmatrix} -P_i^\top A_i P_i + \beta_i I_n & -P_i^\top A_i \big(z + a^i\big) \\ -\big(z + a^i\big)^\top A_i^\top P_i & \zeta^i \end{bmatrix} \succeq O, \\
& \begin{bmatrix} \dfrac{1 - \zeta^i - w_i - \beta_i}{2} \\ \dfrac{1 + \zeta^i + w_i + \beta_i}{2} \\ M_i^\top \big(z + a^i\big) \end{bmatrix} \in \mathcal{K}_{n+2}, \\
& \beta_i \geq 0, \quad i = 1, \ldots, m.
\end{aligned}
\tag{4.23}
$$

Note that the second-order cone constraints can be converted further into semidefinite constraints.

4.6 Numerical experiments

In this section, we present some numerical results using Algorithm 4.1 for the problems in Section 4.5. The experiments are carried out on a machine with a 1.8GHz Intel Core i5 CPU and 8GB memory, and we implement all codes in MATLAB R2017a. We consider the problem (1.1), where $n = 5, m = 2, f_i(x) = (1/2)\langle x, A_i x \rangle + \langle a^i, x \rangle, g_i(x) = \max_{u \in \mathcal{U}_i} \hat{g}_i(x, u), A_i \in \mathbf{R}^{n \times n}, a^i \in \mathbf{R}^n$, and $\hat{g}_i \colon \mathbf{R}^n \to \mathbf{R}, i = 1, \ldots, m$. Here, we assume that each A_i is positive semidefinite so that it can be decomposed as $A_i = M_i M_i^\top$, where $M_i \in \mathbf{R}^{n \times n}$. We generate M_i and a^i by choosing every component randomly from the standard normal distribution. To implement Algorithm 4.1, we make the following choices.

Remark 4.3

- *Every component of x^0 is chosen randomly from the standard normal distribution.*

- *In Experiments 1 and 3, we set the constant $\ell = 5$. In Experiment 2, we set*

the constant $\ell = 7$.

- *The stopping criterion is replaced by* $\|d^k\| < \varepsilon := 10^{-6}$.

Also, we run each one of the following experiments 100 times from different initial points and with $\delta = 0, 0.05, 0.1$. Naturally, when $\delta = 0$, no uncertainties are considered.

Experiment 1

In the first experiment, we solve the problem of Section 4.5.1. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} \langle u, x \rangle, i = 1, 2$, where $\mathcal{U}_1 = \{u \in \mathbf{R}^5 \mid -\delta \leq u_i \leq \delta, i = 1, \ldots, 5\}$ and $\mathcal{U}_2 = \{u \in \mathbf{R}^5 \mid -\delta \leq (Bu)_i \leq \delta, i = 1, \ldots, 5\}$. Here, every component of $B \in \mathbf{R}^{5\times 5}$ is chosen randomly from the standard normal distribution and $\delta \geq 0$. We use the MATLAB solver *linprog* to solve (4.17) and *quadprog* to solve (4.18). Figure 4.1 is the result for this experiment. For each δ, we obtained part of the Pareto frontier, and as δ gets smaller, the objective values become smaller.

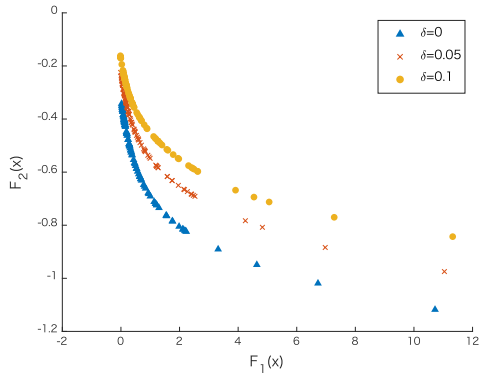


Figure 4.1: Result for Experiment 1

Experiment 2

In the second experiment, we solve the problem of Section 4.5.2. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} \langle u, x \rangle$, where $\mathcal{U}_i = \{u \in \mathbf{R}^5 \mid \|u\|_2 \leq \delta\}, i = 1, 2$. We use the MATLAB solver *SeDuMi* [**Sturm1999**] to solve (4.19). Figure 4.2 is the result of this experiment. Once again, we obtained part of the Pareto frontier for the problems with and without uncertainties.

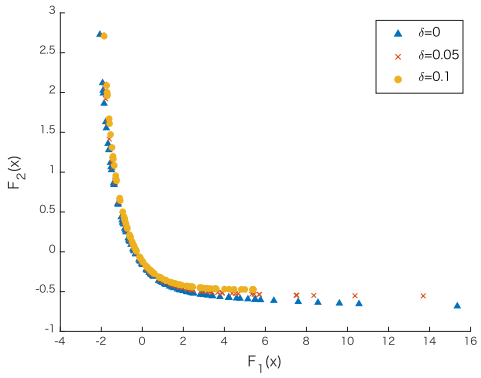


Figure 4.2: Result for Experiment 2

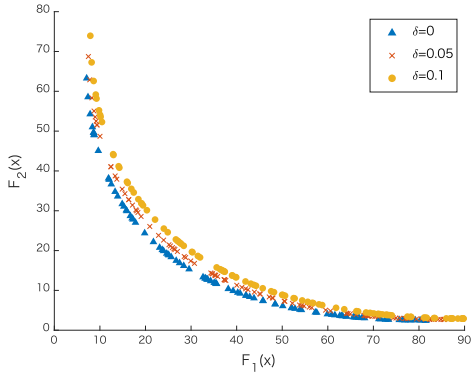


Figure 4.3: Result for Experiment 3

Experiment 3

Now, in the last experiment, we solve the problem of Section 4.5.3. We assume that $g_i(x) = \max_{u \in \mathcal{U}_i} \langle u + x, BB^\top (u + x) \rangle$, where $\mathcal{U}_i = \{u \in \mathbf{R}^5 \mid \|u\| \leq \delta\}, i = 1, 2$. Once again, every component of $B \in \mathbf{R}^{5\times 5}$ is randomly chosen from the standard normal distribution and $\delta \geq 0$. We use the MATLAB solver *fmincon* to solve (4.20) and *SeDuMi* to solve (4.23). As it can be seen in Figure 4.3, we also obtained the Pareto frontier in this case.

4.7 Conclusions

We proposed the proximal gradient method for composite multi-objective optimization problems. Under reasonable assumptions, we proved its global convergence and convergence rate. Moreover, we presented some applications for robust multi-objective optimization. We can convert the subproblems to well-known convex optimization problems in some robust optimization problems. Finally, we carried out some numerical experiments for robust multi-objective optimization problems, and we observed that the Pareto frontier changes when the uncertainty set is modified.

Chapter 5

An accelerated proximal gradient method for multi-objective optimization

5.1 Introduction

This chapter develops the accelerated proximal gradient method for the unconstrained convex composite multi-objective optimization, i.e., (1.14) with f_i being convex and $C = \mathbf{R}^n$.

There are many studies related to the acceleration of single-objective first-order methods. After being established by Nesterov [**Nesterov1983**], researchers developed various accelerated schemes. In particular, the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [**Beck2009**], an accelerated version of the proximal gradient method, has contributed to a wide range of research fields, including image and signal processing. However, the studies associated with accelerated algorithms are still insufficient in the multi-objective case. In 2020, El Moudden and El Mouatasim [**ElMoudden2020**] proposed an accelerated diagonal steepest descent method for multi-objective optimization, a natural extension of Nesterov's accelerated method for single-objective problems. They proved the global convergence rate of the algorithm ($O(1/k^2)$) under the assumption that the sequence of the Lagrange multipliers of the subproblems is eventually fixed. Nevertheless, this assumption is restrictive because it indicates that the approach is essentially the same as Nesterov's (single-objective) method, only applied to the minimization of a weighted

sum of the objective functions.

Here, we propose a genuine accelerated proximal gradient method for multi-objective optimization. As usual, we solve a convex (scalar-valued) subproblem in each iteration. While the accelerated and non-accelerated algorithms solve the same subproblem in the single-objective case, the subproblem of our accelerated method has terms that are not included in the non-accelerated version. However, we can ignore these terms in the single-objective case, and thus we can regard our proposed method as a generalization of FISTA. Moreover, under more natural assumptions, we prove the proposed method's global convergence rate ($O(1/k^2)$) by using a merit function (3.1) to measure the complexity.

The outline of this chapter is as follows. We present the accelerated proximal gradient method for multi-objective optimization in Section 5.2 and analyze its $O(1/k^2)$ convergence rate in Section 5.3. Moreover, Section 5.4 demonstrates the convergence of the iterates. Finally, we report some numerical results for test problems in Section 5.5, demonstrating that the proposed method is faster than the one without acceleration.

5.2 The algorithm

This section proposes an accelerated version of the proximal gradient method for multi-objective optimization. Similar to the non-accelerated version given in the last section, a subproblem is considered in each iteration. More specifically, the proposed method solves the following subproblem for given $x \in \mathrm{dom}(F)$, $y \in \mathbf{R}^n$, and $\alpha > 0$:

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\alpha^{\mathrm{acc}}(z; x, y), \tag{5.1}$$

where

$$\varphi_\alpha^{\mathrm{acc}}(z; x, y) := \max_{i=1,\ldots,m} [\langle \nabla f_i(y), z - y \rangle + g_i(z) + f_i(y) - F_i(x)] + \frac{1}{2\alpha}\|z - y\|_2^2.$$

Note that when $y = x$, (5.1) is reduced to the subproblem (4.1) of the proximal gradient method. Note also that when $m = 1$, the subproblem becomes

$$\min_{z \in \mathbf{R}^n} \quad \langle \nabla f_1(y), z - y \rangle + g_1(z) + \frac{1}{2\alpha}\|z - y\|_2^2, \tag{5.2}$$

which is the subproblem of the single-objective FISTA (Algorithm 1.2). The distinctive feature of our proposal (5.1) is the term $f_i(y) - f_i(x)$, whereas the easy analogy from the single-objective subproblem (5.2) is

$$\min_{z \in \mathbf{R}^n} \quad \max_{i=1,\dots,m} [\langle \nabla f_i(y), z - y \rangle + g_i(z)] + \frac{1}{2\alpha} \|z - y\|_2^2.$$

By putting such a term, the inside of the max operator approximates $F_i(z) - F_i(x)$ rather than $F_i(z) - F_i(y)$. This is a negligible difference in the single-objective case but profoundly affects the proof in the multi-objective case.

Since g_i is convex for all $i = 1, \dots, m$, $z \mapsto \varphi_\alpha^{\mathrm{acc}}(z; x, y)$ is strongly convex. Thus, the subproblem (5.1) has a unique optimal solution $p_\alpha^{\mathrm{acc}}(x, y)$ and takes the optimal function value $\theta_\alpha^{\mathrm{acc}}(x, y)$, i.e.,

$$p_\alpha^{\mathrm{acc}}(x, y) \coloneqq \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi_\alpha^{\mathrm{acc}}(z; x, y) \quad \text{and} \quad \theta_\alpha^{\mathrm{acc}}(x, y) \coloneqq \min_{z \in \mathbf{R}^n} \varphi_\alpha^{\mathrm{acc}}(z; x, y). \tag{5.3}$$

Moreover, the optimality condition of (5.1) implies

$$\frac{1}{\alpha}[y - p_\alpha^{\mathrm{acc}}(x, y)] \in \operatorname*{conv}_{i \in \mathcal{I}_\alpha(x,y)} [\nabla f_i(y) + \partial g_i(p_\alpha^{\mathrm{acc}}(x, y))] \quad \text{for all } x, y \in \mathbf{R}^n$$

with

$$\mathcal{I}_\alpha(x, y) \coloneqq \operatorname*{argmax}_{i=1,\dots,m} [\langle \nabla f_i(y), p_\alpha^{\mathrm{acc}}(x, y) - y \rangle + g_i(p_\alpha^{\mathrm{acc}}(x, y)) + f_i(y) - F_i(x)]. \tag{5.4}$$

Hence, for any $x, y, z \in \mathbf{R}^n$, there exists $\lambda(x, y) \in \Delta^m$ such that $\lambda_j(x, y) = 0$ for $j \notin \mathcal{I}_\alpha(x, y)$ and

$$\begin{aligned} &\frac{1}{\alpha} \langle y - p_\alpha^{\mathrm{acc}}(x, y), z - p_\ell^{\mathrm{acc}}(x, y) \rangle \\ &\qquad \le \sum_{i=1}^m \lambda_i(x, y) [\langle \nabla f_i(y), z - p_\alpha^{\mathrm{acc}}(x, y) \rangle + g_i(z) - g_i(p_\alpha^{\mathrm{acc}}(x, y))], \end{aligned} \tag{5.5}$$

where Δ^m denotes the unit m-simplex (2.1). We also note that by taking $z = y$ in the objective function of (5.1), we have

$$\theta_\alpha^{\mathrm{acc}}(x, y) \le \varphi_\alpha^{\mathrm{acc}}(y; x, y) = \max_{i=1,\dots,m} [F_i(y) - F_i(x)] \tag{5.6}$$

for all $x \in \mathrm{dom}(F)$ and $y \in \mathbf{R}^n$. We now characterize weak Pareto optimality in terms of the mappings p_α^{acc} and $\theta_\alpha^{\mathrm{acc}}$, similarly to Lemma 4.1 for the proximal gradient method.

Proposition 5.1
Let $p_\alpha^{\mathrm{acc}}(x,y)$ and $\theta_\alpha^{\mathrm{acc}}(x,y)$ be defined by (5.3). Then, the statements below hold.

(i) The following three conditions are equivalent:

(a) $y \in \mathbf{R}^n$ is weakly Pareto optimal for (1.1);

(b) $p_\alpha^{\mathrm{acc}}(x,y) = y$ for some $x \in \mathbf{R}^n$;

(c) $\theta_\alpha^{\mathrm{acc}}(x,y) = \max_{i=1,\dots,m}[F_i(y) - F_i(x)]$ for some $x \in \mathbf{R}^n$.

(ii) The mappings p_α^{acc} and $\theta_\alpha^{\mathrm{acc}}$ are continuous. Particularly, if ∇f_i is locally Lipschitz continuous, p_α^{acc} and $\theta_\alpha^{\mathrm{acc}}$ are locally Hölder continuous with exponent $1/2$ and locally Lipschitz continuous, respectively.

Proof. Claim (i): From (5.6) and the fact that $\theta_\alpha^{\mathrm{acc}}(x,y) = \varphi_\alpha^{\mathrm{acc}}(p_\alpha^{\mathrm{acc}}(x,y);x,y)$, the equivalence between (b) and (c) is apparent. Let us show that (a) and (b) are equivalent. When y is weakly Pareto optimal, we can immediately see from Lemma 4.1 that $p_\alpha^{\mathrm{acc}}(x,y) = p_\alpha(y) = y$ by letting $x = y$. Conversely, suppose that $p_\alpha^{\mathrm{acc}}(x,y) = y$ for some $x \in \mathbf{R}^n$. Let $z \in \mathbf{R}^n$ and $\beta \in (0,1)$. The optimality of $p_\alpha^{\mathrm{acc}}(x,y) = y$ for (5.1) gives

$$\begin{aligned}\max_{i=1,\dots,m}[F_i(y) - F_i(x)] &\le \varphi_\alpha^{\mathrm{acc}}(y + \beta(z-y);x,y)\\ &= \max_{i=1,\dots,m}[\langle \nabla f_i(y), \beta(z-y)\rangle + g_i(y+\beta(z-y)) + f_i(y) - F_i(x)]\\ &\qquad + \frac{1}{2\alpha}\|\beta(z-y)\|_2^2.\end{aligned}$$

Thus, from the convexity of f_i, we get

$$\max_{i=1,\dots,m}[F_i(y) - F_i(x)] \le \max_{i=1,\dots,m}[F_i(y+\beta(z-y)) - F_i(x)] + \frac{1}{2\alpha}\|\beta(z-y)\|_2^2.$$

Moreover, the convexity of F_i yields

$$\begin{aligned}
&\max_{i=1,\dots,m}[F_i(y)-F_i(x)]\\
&\qquad\le \max_{i=1,\dots,m}[\beta F_i(z)+(1-\beta)F_i(y)-F_i(x)]+\frac{1}{2\alpha}\|\beta(z-y)\|_2^2\\
&\qquad\le \beta\max_{i=1,\dots,m}[F_i(z)-F_i(y)]+\max_{i=1,\dots,m}[F_i(y)-F_i(x)]+\frac{1}{2\alpha}\|\beta(z-y)\|_2^2.
\end{aligned}$$

Therefore, we get

$$\max_{i=1,\dots,m}[F_i(z)-F_i(y)]\ge-\frac{\beta}{2\alpha}\|z-y\|_2^2.$$

Taking $\beta\searrow 0$, we obtain

$$\max_{i=1,\dots,m}[F_i(z)-F_i(y)]\ge 0,$$

which implies the weak Pareto optimality of y.

Claim (ii): Let Ω be a bounded subset of $\mathbf{R}^n$ and take $\hat{x},\hat{y},\check{x},\check{y}\in\Omega$. Adding (5.5) with $(x,y,z):=(\hat{x},\hat{y},p_\alpha^{\mathrm{acc}}(\check{x},\check{y})),(\check{x},\check{y},p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}))$ gives

$$\begin{aligned}
&\frac{1}{\alpha}\langle p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})-p_\alpha^{\mathrm{acc}}(\check{x},\check{y})-(\hat{y}-\check{y}),p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})-p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle\\
&\quad\le\sum_{i=1}^m\lambda_i(\hat{x},\hat{y})[\langle\nabla f_i(\hat{y}),p_\alpha^{\mathrm{acc}}(\check{x},\check{y})-p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})\rangle+g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y}))-g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}))]\\
&\qquad+\sum_{i=1}^m\lambda_i(\check{x},\check{y})[\langle\nabla f_i(\check{y}),p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})-p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle+g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}))-g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y}))]\\
&\quad\le\sum_{i=1}^m\lambda_i(\hat{x},\hat{y})[\langle\nabla f_i(\hat{y}),\hat{y}-p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})\rangle-g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}))-f_i(\hat{y})+F_i(\hat{x})]\\
&\qquad+\sum_{i=1}^m\lambda_i(\check{x},\check{y})[\langle\nabla f_i(\check{y}),\check{y}-p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle-g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y}))-f_i(\check{y})+F_i(\check{x})]\\
&\qquad+\sum_{i=1}^m\lambda_i(\hat{x},\hat{y})[\langle\nabla f_i(\hat{y}),p_\alpha^{\mathrm{acc}}(\check{x},\check{y})-\hat{y}\rangle+g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y}))+f_i(\hat{y})-F_i(\hat{x})]\\
&\qquad+\sum_{i=1}^m\lambda_i(\check{x},\check{y})[\langle\nabla f_i(\check{y}),p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})-\check{y}\rangle+g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}))+f_i(\check{y})-F_i(\check{x})].
\end{aligned}$$

Since $\lambda_j(x,y)$ for $j \in \mathcal{I}_\alpha(x)$ with $\mathcal{I}_\alpha$ given by (5.4), we get

$$\begin{aligned}
&\frac{1}{\alpha}\langle p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - (\hat{y}-\check{y}), p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle \\
&\quad \leq \min_{i=1,\dots,m}[\langle \nabla f_i(\hat{y}), \hat{y} - p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})\rangle - g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})) - f_i(\hat{y}) + F_i(\hat{x})] \\
&\qquad + \min_{i=1,\dots,m}[\langle \nabla f_i(\check{y}), \check{y} - p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle - g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) - f_i(\check{y}) + F_i(\check{x})] \\
&\qquad + \sum_{i=1}^m \lambda_i(\hat{x},\hat{y})[\langle \nabla f_i(\hat{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\hat{y}) - F_i(\hat{x})] \\
&\qquad + \sum_{i=1}^m \lambda_i(\check{x},\check{y})[\langle \nabla f_i(\check{y}), p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - \check{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})) + f_i(\check{y}) - F_i(\check{x})] \\
&\quad \leq \sum_{i=1}^m \lambda_i(\check{x},\check{y})[\langle \nabla f_i(\hat{y}), \hat{y} - p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})\rangle - g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})) - f_i(\hat{y}) + F_i(\hat{x})] \\
&\qquad + \sum_{i=1}^m \lambda_i(\hat{x},\hat{y})[\langle \nabla f_i(\check{y}), \check{y} - p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle - g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) - f_i(\check{y}) + F_i(\check{x})] \\
&\qquad + \sum_{i=1}^m \lambda_i(\hat{x},\hat{y})[\langle \nabla f_i(\hat{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\hat{y}) - F_i(\hat{x})] \\
&\qquad + \sum_{i=1}^m \lambda_i(\check{x},\check{y})[\langle \nabla f_i(\check{y}), p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - \check{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})) + f_i(\check{y}) - F_i(\check{x})].
\end{aligned}$$

Thus, quick calculations imply

$$\begin{aligned}
&\frac{1}{\alpha}\|p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\|_2^2 \leq \frac{1}{\alpha}\langle p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - p_\alpha^{\mathrm{acc}}(\check{x},\check{y}), \hat{y}-\check{y}\rangle \\
&\quad \leq \sum_{i=1}^m \lambda_i(\hat{x},\hat{y})[\langle \nabla f_i(\hat{y}) - \nabla f_i(\check{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\rangle + f_i(\hat{y}) - f_i(\check{y}) - (F_i(\hat{x}) - F_i(\check{x}))] \\
&\qquad + \sum_{i=1}^m \lambda_i(\check{x},\check{y})[\langle \nabla f_i(\check{y}) - \nabla f_i(\hat{y}), p_\alpha^{\mathrm{acc}}(\hat{x},\hat{y})\rangle + f_i(\check{y}) - f_i(\hat{x}) - (F_i(\check{x}) - F_i(\hat{x}))] \\
&\qquad + \sum_{i=1}^m [\lambda_i(\check{x},\check{y}) - \lambda_i(\hat{x},\hat{y})][\langle \nabla f_i(\hat{y}) - \nabla f_i(\check{y}), \hat{y}\rangle + \langle \nabla f_i(\check{y}), \hat{y}-\check{y}\rangle].
\end{aligned}$$

When $(\hat{x},\hat{y}) \to (\check{x},\check{y})$, the right-hand side tends to zero, so p_α^{acc} and $\theta_\alpha^{\mathrm{acc}}$ are continuous. Moreover, assume that ∇f_i is locally Lipschitz continuous for $i = 1,\dots,m$. Since f_i, g_i, F_i are also locally Lipschitz, the above inequality shows that p_α^{acc} is locally Hölder continuous with exponent $1/2$.

On the other hand, the definition (5.3) of p_α^{acc} and $\theta_\alpha^{\mathrm{acc}}$ gives

$$\begin{aligned}
&\theta_\alpha^{\mathrm{acc}}(\hat{x},\hat{y}) - \theta_\alpha^{\mathrm{acc}}(\check{x},\check{y}) \leq \varphi_\alpha^{\mathrm{acc}}(p_\alpha^{\mathrm{acc}}(\check{x},\check{y});\hat{x},\hat{y}) - \varphi_\alpha^{\mathrm{acc}}(p_\alpha^{\mathrm{acc}}(\check{x},\check{y});\check{x},\check{y}) \\
&\quad = \max_{i=1,\dots,m}[\langle \nabla f_i(\hat{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\hat{y}) - F_i(\hat{x})] \\
&\qquad - \max_{i=1,\dots,m}[\langle \nabla f_i(\check{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \check{y}\rangle + g_i(p_\alpha^{\mathrm{acc}}(\check{x},\check{y})) + f_i(\check{y}) - F_i(\check{x})] \\
&\qquad + \frac{1}{2\alpha}\left[\|p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\|_2^2 - \|p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \check{y}\|_2^2\right] \\
&\quad \leq \max_{i=1,\dots,m}[\langle \nabla f_i(\check{y}), \check{y} - \hat{y}\rangle + \langle \nabla f_i(\hat{y}) - \nabla f_i(\check{y}), p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y}\rangle \\
&\qquad\qquad + f_i(\hat{y}) - f_i(\check{y}) - F_i(\hat{x}) + F_i(\check{x})] \\
&\qquad + \frac{1}{2\alpha}\langle 2p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y} - \check{y}, \check{y} - \hat{y}\rangle \\
&\quad \leq \max_{i=1,\dots,m}\|\nabla f_i(\check{y})\|_2\|\hat{y} - \check{y}\|_2 + \|\hat{y} - p_\alpha^{\mathrm{acc}}(\check{x},\check{y})\|_2 \max_{i=1,\dots,m}\|\nabla f_i(\hat{y}) - \nabla f_i(\check{y})\|_2 \\
&\qquad + \max_{i=1,\dots,m}|f_i(\hat{y}) - f_i(\check{y})| + \max_{i=1,\dots,m}|F_i(\hat{x}) - F_i(\check{x})| \\
&\qquad + \frac{1}{2\alpha}\|2p_\alpha^{\mathrm{acc}}(\check{x},\check{y}) - \hat{y} - \check{y}\|_2\|\hat{y} - \check{y}\|_2,
\end{aligned}$$

where the second inequality follows from (2.2), and the third inequality comes from the Cauchy-Schwarz inequalities. Since the above inequality holds even if we interchange $(\hat{x},\hat{y})$ and $(\check{x},\check{y})$, we can show the Lipschitz continuity of $\theta_\alpha^{\mathrm{acc}}$ on Ω in the same way as in the previous paragraph. □

Proposition 5.1 suggests that we can use $\|p_\alpha^{\mathrm{acc}}(x,y) - y\|_\infty < \varepsilon$ for some $\varepsilon > 0$ as a stopping criterion. Now, we state below the proposed algorithm.

The sequence $\{t_k\}$ defined in lines 2 and 8 of Algorithm 5.1 generalizes the well-known momentum factors in single-objective accelerated methods. For example, when $a = 0$ and $b = 1/4$, they coincide with the one in Algorithm 5.1 and the original FISTA [**Nesterov1983**, **Beck2009**] ($t_1 = 1$ and $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$). Moreover, if $b = a^2/4$, then $\{t_k\}$ has the general term $t_k = (1-a)k/2 + (1+a)/2$, which corresponds to the one used in [**Chambolle2015**, **Su2016**, **Attouch2016**, **Attouch2018**]. This means that our generalization allows finer algorithm tuning by varying a and b. We show below some properties of $\{t_k\}$ and $\{\gamma_k\}$.

Lemma 5.1

Let $\{t_k\}$ and $\{\gamma_k\}$ be defined by lines 2, 8 and 9 in Algorithm 5.1 for arbitrary $a \in [0,1)$ and $b \in [a^2/4, 1/4]$. Then, the following inequalities hold for all $k \geq 1$.

(i) $t_{k+1} \geq t_k + \dfrac{1-a}{2}$ *and* $t_k \geq \dfrac{1-a}{2}k + \dfrac{1+a}{2}$;

Algorithm 5.1 Accelerated proximal gradient method with general stepsizes for (1.1)

Input: $x^0 = y^1 \in \mathrm{dom}(F), a \in [0,1), b \in [a^2/4, 1/4], \varepsilon > 0$.
Output: x^*: A weakly Pareto optimal point

1: $k \leftarrow 1$
2: $t_1 \leftarrow 1$
3: **loop**
4: $x^k \leftarrow p_\alpha^{\mathrm{acc}}(x^{k-1}, y^k)$ for some stepsize $\alpha > 0$
5: **if** $\|x^k - y^k\|_\infty < \varepsilon$ **then**
6: **return** x^k
7: **end if**
8: $t_{k+1} \leftarrow \sqrt{t_k^2 - at_k + b} + 1/2$
9: $\gamma_k \leftarrow (t_k - 1)/t_{k+1}$
10: $y^{k+1} \leftarrow x^k + \gamma_k(x^k - x^{k-1})$
11: $k \leftarrow k + 1$
12: **end loop**

(ii) $t_{k+1} \le t_k + \dfrac{1 - a + \sqrt{4b - a^2}}{2}$ *and* $t_k \le \dfrac{1 - a + \sqrt{4b - a^2}}{2}(k-1) + 1 \le k$;

(iii) $t_k^2 - t_{k+1}^2 + t_{k+1} = at_k - b + \dfrac{1}{4} \ge at_k$;

(iv) $0 \le \gamma_k \le \dfrac{k-1}{k+1/2}$;

(v) $1 - \gamma_k^2 \ge \dfrac{1}{t_k}$.

Proof. Claim (i): From the definition of $\{t_k\}$, we have

$$t_{k+1} = \sqrt{t_k^2 - at_k + b} + \frac{1}{2} = \sqrt{\left(t_k - \frac{a}{2}\right)^2 + \left(b - \frac{a^2}{4}\right)} + \frac{1}{2}. \tag{5.7}$$

Since $b \ge a^2/4$, we get

$$t_{k+1} \ge \left|t_k - \frac{a}{2}\right| + \frac{1}{2}.$$

Since $t_1 = 1 \ge a/2$, we can quickly see that $t_k \ge a/2$ for any k by induction. Thus, we have

$$t_{k+1} \ge t_k + \frac{1-a}{2}.$$

Applying the above inequality recursively, we obtain

$$t_k \geq \frac{1-a}{2}(k-1) + t_1 = \frac{1-a}{2}k + \frac{1+a}{2}.$$

Claim (ii): From (5.7) and the relation $\sqrt{\beta_1+\beta_2} \leq \sqrt{\beta_1} + \sqrt{\beta_2}$ with $\beta_1, \beta_2 \geq 0$, we get the first inequality. Using it recursively, it follows that

$$t_k \leq \frac{1-a+\sqrt{4b-a^2}}{2}(k-1) + t_1 = \frac{1-a+\sqrt{4b-a^2}}{2}(k-1) + 1.$$

Since $a \in [0,1), b \in [a^2/4, 1/4]$, we observe that

$$\frac{1-a+\sqrt{4b-a^2}}{2} \leq \frac{1-a+\sqrt{1-a^2}}{2} \leq 1.$$

Hence, the above two inequalities lead to the desired result.

Claim (iii): An easy computation shows that

$$\begin{aligned} t_k^2 - t_{k+1}^2 + t_{k+1} &= t_k^2 - \left[\sqrt{t_k^2 - at_k + b} + \frac{1}{2}\right]^2 + \sqrt{t_k^2 - at_k + b} + \frac{1}{2} \\ &= at_k - b + \frac{1}{4} \geq at_k, \end{aligned}$$

where the inequality holds since $b \leq 1/4$.

Claim (iv): The first inequality is clear from the definition of γ_k since claim (i) yields $t_k \geq 1$. Again, the definition of γ_k and claim (i) give

$$\gamma_k = \frac{t_k - 1}{t_{k+1}} \leq \frac{t_k - 1}{t_k + (1-a)/2} = 1 - \frac{3-a}{2t_k + 1 - a}.$$

Combining with the claim (ii), we get

$$\begin{aligned} \gamma_k &\leq 1 - \frac{3-a}{\left(1-a+\sqrt{4b-a^2}\right)(k-1) + 3 - a} \\ &= \frac{\left(1-a+\sqrt{4b-a^2}\right)(k-1)}{\left(1-a+\sqrt{4b-a^2}\right)(k-1) + 3 - a} \\ &= \frac{k-1}{k-1+(3-a)/\left(1-a+\sqrt{4b-a^2}\right)}. \end{aligned} \tag{5.8}$$

On the other hand, it follows that

$$\min_{a\in[0,1),b\in[a^2/4,1/4]} \frac{3-a}{1-a+\sqrt{4b-a^2}} = \min_{a\in[0,1)} \frac{3-a}{1-a+\sqrt{1-a^2}} = \frac{3}{2}, \tag{5.9}$$

where the second equality follows from the monotonic non-decreasing property implied by

$$\frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{3-a}{1-a+\sqrt{1-a^2}}\right) = \frac{2\sqrt{1-a^2}+3a-1}{\left(\sqrt{1-a^2}-a+1\right)^2\sqrt{1-a^2}} > 0 \quad \text{for all } a\in[0,1).$$

Combining (5.8) and (5.9), we obtain $\gamma_k \leq (k-1)/(k+1/2)$.

Claim (v): claim (i) implies that $t_{k+1} > t_k \geq 1$. Thus, the definition of γ_k implies that

$$1-\gamma_k^2 = 1-\left(\frac{t_k-1}{t_{k+1}}\right)^2 \geq 1-\left(\frac{t_k-1}{t_k}\right)^2 = \frac{2t_k-1}{t_k^2} \geq \frac{2t_k-t_k}{t_k^2} = \frac{1}{t_k}.$$

□

We end this section by noting some remarks about the proposed algorithm.

Remark 5.1

(i) *Since* $x \in \mathrm{dom}(F)$ *implies* $p_\alpha^{\mathrm{acc}}(x,y) \in \mathrm{dom}(F)$, *every* x^k *computed by the above algorithm is in* $\mathrm{dom}(F)$. *However,* y^k *is not necessarily in* $\mathrm{dom}(F)$.

(ii) *Since* $y^1 = x^0$, *it follows from* (5.6) *that*

$$\theta_\alpha^{\mathrm{acc}}(x^0,y^1) \leq 0,$$

but the inequality $\theta_\alpha^{\mathrm{acc}}(x^{k-1},y^k) \leq 0$ *does not necessarily hold for* $k \geq 2$.

(iii) *When* $m=1$, *we can remove the term* $f_i(y)-F_i(x)$ *from the subproblem* (5.1), *so Algorithm 5.1 corresponds to the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)* [**Beck2009**] *for single-objective optimization.*

(iv) *Algorithm 5.1 induces the accelerated versions of first-order algorithms such as the steepest descent* [**Fliege2000**], *proximal point* [**Bonnel2005**], *and projected gradient methods* [**Grana-Drummond2004**].

5.3 Convergence rates analysis

This section shows that Algorithm 5.1 has the $O(1/k^2)$ convergence rate under Assumption 4.2. For simplicity, we suppose the constant stepsize $\alpha \in (0, 1/L_f]$. Then, Lemmas 2.2 and 2.3 implies

$$\theta_\alpha^{\mathrm{acc}}(x, y) \geq \max_{i=1,\dots,m} [F_i(p_\alpha^{\mathrm{acc}}(x, y)) - F_i(x)] \tag{5.10}$$

for all $x \in \mathrm{dom}(F)$ and $y \in \mathbf{R}^n$. When it is challenging to estimate L_f, we can use the backtracking procedure like Section 4.2.2.

We present below the main theorem of this section.

Theorem 5.1

Let $\{x^k\}$ be a sequence generated by Algorithm 5.1 and recall that u_∞ is given by (3.1). Then, the following two equations hold:

(i) $F_i(x^k) \leq F_i(x^0)$ *for all* $i = 1, \dots, m$ *and* $k \geq 0$;

(ii) $u_\infty(x^k) = O(1/k^2)$ *as* $k \to \infty$ *under Assumption 4.4.*

Claim (i) means that $\{x^k\} \subseteq \mathbf{lev}_F(F(x^0))$, where $\mathbf{lev}_{F(x^0)}(F)$ denotes the level set of F (cf. (2.5)). Note, however, that the objective functions are generally not monotonically non-increasing. Claim (ii) also claims the global convergence rate.

Before proving Theorem 5.1, let us give several lemmas. First, we present some properties of $\{t_k\}$ and $\{\gamma_k\}$. As in [**Tanabe2022a**], we also introduce $\sigma_k \colon \mathbf{R}^n \to [-\infty, +\infty)$ and $\rho_k \colon \mathbf{R}^n \to \mathbf{R}$ for $k \geq 0$ as follows, which assist the analysis:

$$\begin{aligned} \sigma_k(z) &:= \min_{i=1,\dots,m} [F_i(x^k) - F_i(z)], \\ \rho_k(z) &:= \|t_{k+1}x^{k+1} - (t_{k+1} - 1)x^k - z\|_2^2. \end{aligned} \tag{5.11}$$

The following lemma on σ_k is helpful in the subsequent discussions.

Lemma 5.2

Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Then, the following inequalities hold for all $z \in \mathbf{R}^n$ and $k \geq 0$:

(i) $\sigma_{k+1}(z) \leq -\dfrac{1}{\alpha}\langle y^{k+1} - x^{k+1}, z - y^{k+1}\rangle - \dfrac{1}{2\alpha}\|x^{k+1} - y^{k+1}\|_2^2$;

(ii) $\sigma_{k+1}(z) - \sigma_k(z) \leq \max_{i=1,\dots,m} [F_i(x^{k+1}) - F_i(x^k)] \leq -\dfrac{1}{\alpha}\langle y^{k+1} - x^{k+1}, x^k - y^{k+1}\rangle - \dfrac{1}{2\alpha}\|x^{k+1} - y^{k+1}\|_2^2$.

Proof. Suppose that $z \in \mathbf{R}^n$ and $k \geq 0$. Recall that a Lagrange multiplier $\lambda\big(x^k, y^{k+1}\big) \in \Delta^m$ exists and satisfies the optimality condition (5.5).

Claim (i): From the definition (5.11) of σ_{k+1}, we get

$$\begin{aligned}\sigma_{k+1}(z) &= \min_{i=1,\dots,m}\big[F_i\big(x^{k+1}\big) - F_i(z)\big] \\ &\leq \sum_{i=1}^{m} \lambda_i\big(x^k, y^{k+1}\big)\big[F_i\big(x^{k+1}\big) - F_i(z)\big] \\ &= \sum_{i=1}^{m} \lambda_i\big(x^k, y^{k+1}\big)\big[F_i\big(x^{k+1}\big) - F_i\big(x^k\big) + F_i\big(x^k\big) - F_i(z)\big].\end{aligned}$$

where the inequality follows from $\lambda\big(x^k, y^{k+1}\big) \in \Delta^m$. From (5.10), we have

$$\begin{aligned}&\sigma_{k+1}(z) \leq \theta_\alpha^{\mathrm{acc}}\big(x^k, y^{k+1}\big) + \sum_{i=1}^{m} \lambda_i\big(x^k, y^{k+1}\big)\big[F_i\big(x^k\big) - F_i(z)\big] \\ &= \sum_{i=1}^{m} \lambda_i\big(x^k, y^{k+1}\big)\big[\big\langle \nabla f_i\big(y^{k+1}\big), x^{k+1} - y^{k+1}\big\rangle + g_i\big(x^{k+1}\big) + f_i\big(y^{k+1}\big) - F_i(z)\big] \\ &\quad + \frac{1}{2\alpha}\big\|x^{k+1} - y^{k+1}\big\|_2^2,\end{aligned}$$

where the second inequality comes from (5.3) and (5.4). The convexity of f_i yields

$$\begin{aligned}&\sigma_{k+1}(z) \\ &\leq \sum_{i=1}^{m} \lambda_i\big(x^k, y^{k+1}\big)\big[\big\langle \nabla f_i\big(y^{k+1}\big), x^{k+1} - z\big\rangle + g_i\big(x^{k+1}\big) - g_i(z)\big] + \frac{1}{2\alpha}\big\|x^{k+1} - y^{k+1}\big\|_2^2.\end{aligned}$$

Equation (5.5) gives

$$\begin{aligned}\sigma_{k+1}(z) &\leq -\frac{1}{\alpha}\big\langle y^{k+1} - x^{k+1}, z - x^{k+1}\big\rangle + \frac{1}{2\alpha}\big\|x^{k+1} - y^{k+1}\big\|_2^2 \\ &= -\frac{1}{\alpha}\big\langle y^{k+1} - x^{k+1}, z - y^{k+1}\big\rangle - \frac{1}{2\alpha}\big\|x^{k+1} - y^{k+1}\big\|_2^2.\end{aligned}$$

Claim (ii): Again, from the definition (5.11) of σ_k, we obtain

$$\begin{aligned}
&\sigma_{k+1}(z) - \sigma_k(z) \\
&\quad = \min_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i(z)\right] - \min_{i=1,\dots,m}\left[F_i\left(x^{k}\right) - F_i(z)\right] \\
&\quad \leq \max_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i\left(x^{k}\right)\right],
\end{aligned}$$

where the inequality holds because of (2.2). Equation (5.10) leads to

$$\begin{aligned}
&\max_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i\left(x^{k}\right)\right] \leq \theta_\alpha^{\mathrm{acc}}\left(x^k, y^{k+1}\right) \\
&= \sum_{i=1}^{m} \lambda_i\left(x^k, y^{k+1}\right)\left[\left\langle \nabla f_i(y^{k+1}), x^{k+1} - y^{k+1}\right\rangle + g_i\left(x^{k+1}\right) + f_i\left(y^{k+1}\right) - F_i\left(x^k\right)\right] \\
&\quad + \frac{1}{2\alpha}\left\|x^{k+1} - y^{k+1}\right\|_2^2
\end{aligned}$$

where the equality is from (5.3) and (5.4). From the convexity of f_i, we obtain

$$\begin{aligned}
&\max_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i\left(x^{k}\right)\right] \\
&\quad \leq \sum_{i=1}^{m} \lambda_i\left(x^k, y^{k+1}\right)\left[\left\langle \nabla f_i(y^{k+1}), x^{k+1} - x^{k}\right\rangle + g_i\left(x^{k+1}\right) - g_i\left(x^{k}\right)\right] \\
&\quad + \frac{1}{2\alpha}\left\|x^{k+1} - y^{k+1}\right\|_2^2.
\end{aligned}$$

Equation (5.5) yields

$$\begin{aligned}
\max_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i\left(x^{k}\right)\right] &\leq -\frac{1}{\alpha}\left\langle y^{k+1} - x^{k+1}, x^k - x^{k+1}\right\rangle + \frac{1}{2\alpha}\left\|x^{k+1} - y^{k+1}\right\|_2^2 \\
&= -\frac{1}{\alpha}\left\langle y^{k+1} - x^{k+1}, x^k - y^{k+1}\right\rangle - \frac{1}{2\alpha}\left\|x^{k+1} - y^{k+1}\right\|_2^2.
\end{aligned}$$

□

Therefore, from Lemma 5.1 (v), we can obtain the following result quickly in the same way as in the proof of [**Tanabe2022a**].

Lemma 5.3
Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Then, we have

$$\begin{aligned}\sigma_{k_2}(z) - \sigma_{k_1}(z) &\leq \max_{i=1,\dots,m}\left[F_i\left(x^{k_2}\right) - F_i\left(x^{k_1}\right)\right] \\ &\leq -\frac{1}{2\alpha}\left(\left\|x^{k_2} - x^{k_2-1}\right\|_2^2 - \left\|x^{k_1} - x^{k_1-1}\right\|_2^2 + \sum_{k=k_1}^{k_2-1}\frac{1}{t_k}\left\|x^k - x^{k-1}\right\|_2^2\right)\end{aligned}$$

for any $k_2 \geq k_1 \geq 1$.

Proof. Let $k \geq 1$. From Lemma 5.2 (ii) is equivalent to

$$\begin{aligned}\sigma_{k+1}(z) - \sigma_k(z) &\leq \max_{i=1,\dots,m}\left[F_i\left(x^{k+1}\right) - F_i\left(x^k\right)\right] \leq -\frac{1}{2\alpha}\left(\left\|x^{k+1} - x^k\right\|_2^2 - \left\|y^{k+1} - x^k\right\|_2^2\right) \\ &= -\frac{1}{2\alpha}\left(\left\|x^{k+1} - x^k\right\|_2^2 - \gamma_k^2\left\|x^k - x^{k-1}\right\|_2^2\right),\end{aligned}$$

where the equality comes from line 10 of Algorithm 5.1. Adding up this inequality from $k = k_1$ to $k = k_2 - 1$ yields

$$\begin{aligned}\sigma_{k_2}(z) - \sigma_{k_1}(z) &\leq \max_{i=1,\dots,m}\left[F_i\left(x^{k_2}\right) - F_i\left(x^{k_1}\right)\right] \\ &\leq -\frac{1}{2\alpha}\left[\left\|x^{k_2} - x^{k_2-1}\right\|_2^2 - \left\|x^{k_1} - x^{k_1-1}\right\|_2^2 + \sum_{k=k_1}^{k_2-1}\left(1-\gamma_k^2\right)\left\|x^k - x^{k-1}\right\|_2^2\right].\end{aligned}$$

Using Lemma 5.1 (v), we get the desired inequality. □

We can now show the first part of Theorem 5.1.

Proof of Theorem 5.1 (i). It is clear from Lemma 5.3 with $k_1 = 0$ and $k_2 = k$. □

The next step is to prepare the proof of Theorem 5.1 (ii).

Lemma 5.4
Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 5.1. Also, let σ_k and ρ_k be

defined by (5.11). Then, we have

$$\begin{aligned}\frac{1}{1-a}\Big[t_{k+1}^2 - at_{k+1} + \Big(\frac{1}{4}-b\Big)k\Big]\sigma_{k+1}(z) \\ + \frac{1}{2\alpha(1-a)}\Big[a(t_{k+1}^2 - t_{k+1}) + \Big(\frac{1}{4}-b\Big)k\Big]\big\|x^{k+1}-x^k\big\|_2^2 \\ + \frac{1}{2\alpha(1-a)}\sum_{p=1}^{k}\Big[a^2(t_p-1) + \Big(\frac{1}{4}-b\Big)\frac{p-t_p+a(t_p-1)}{t_p}\Big]\big\|x^p-x^{p-1}\big\|_2^2 \\ + \frac{1}{2\alpha}\rho_k(z) \le \frac{1}{2\alpha}\big\|x^0-z\big\|_2^2.\end{aligned}$$

for all $k \ge 0$ and $z \in \mathbf{R}^n$.

Proof. Let $p \ge 1$ and $z \in \mathbf{R}^n$. Adding Lemma 5.2 (ii) multiplied by $(t_{p+1}-1)$ and Lemma 5.2 (i), both with $k = p$, yields

$$\begin{aligned}t_{p+1}\sigma_{p+1}(z) - (t_{p+1}-1)\sigma_p(z) \\ \le -\frac{1}{2\alpha}\Big[t_{p+1}\big\|x^{p+1}-y^{p+1}\big\|_2^2 + 2\big\langle x^{p+1}-y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1}-1)x^p - z\big\rangle\Big].\end{aligned}$$

Multiplying this inequality by t_{p+1} and using the relation $t_p^2 = t_{p+1}^2 - t_{p+1} + (at_p - b + 1/4)$ (cf. Lemma 5.1 (iii)), we get

$$\begin{aligned}t_{p+1}^2\sigma_{p+1}(z) - t_p^2\sigma_p(z) \le -\frac{1}{2\alpha}\Big[\big\|t_{p+1}(x^{p+1}-y^{p+1})\big\|_2^2 \\ + 2t_{p+1}\big\langle x^{p+1}-y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1}-1)x^p - z\big\rangle\Big] - \Big(at_p - b + \frac{1}{4}\Big)\sigma_p(z).\end{aligned}$$

Applying (2.3) to the right-hand side of the above inequality with

$$v^1 := t_{p+1}y^{p+1}, \quad v^2 := t_{p+1}x^{p+1}, \quad v^3 := (t_{p+1}-1)x^p + z,$$

we get

$$\begin{aligned}t_{p+1}^2\sigma_{p+1}(z) - t_p^2\sigma_p(z) \\ \le -\frac{1}{2\alpha}\Big[\big\|t_{p+1}x^{p+1} - (t_{p+1}-1)x^p - z\big\|_2^2 - \big\|t_{p+1}y^{p+1} - (t_{p+1}-1)x^p - z\big\|_2^2\Big] \\ - \Big(at_p - b + \frac{1}{4}\Big)\sigma_p(z).\end{aligned}$$

Recall that $\rho_p(z) \coloneqq \|t_{p+1}x^{p+1} - (t_{p+1}-1)x^p - z\|_2^2$. Then, considering the definition of y^p given in line 10 of Algorithm 5.1, we obtain

$$t_{p+1}^2\sigma_{p+1}(z) - t_p^2\sigma_p(z) \leq -\frac{1}{2\alpha}[\rho_p(z) - \rho_{p-1}(z)] - \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).$$

Now, let $k \geq 0$. Lemma 5.3 with $(k_1, k_2) = (p, k+1)$ implies

$$t_{p+1}^2\sigma_{p+1}(z) - t_p^2\sigma_p(z) \leq -\frac{1}{2\alpha}[\rho_p(z) - \rho_{p-1}(z)]$$
$$-\left(at_p - b + \frac{1}{4}\right)\left[\sigma_{k+1}(z) + \frac{1}{2\alpha}\left(\left\|x^{k+1} - x^k\right\|_2^2 - \left\|x^p - x^{p-1}\right\|_2^2 + \sum_{r=p}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|_2^2\right)\right].$$

Adding up the above inequality from $p = 1$ to $p = k$, the fact that $t_1 = 1$ and $\rho_0(z) = \|x^1 - z\|_2^2$ leads to

$$\begin{aligned} t_{k+1}^2\sigma_{k+1}(z) - \sigma_1(z) \leq & -\frac{1}{2\alpha}\left[\rho_k(z) - \left\|x^1 - z\right\|_2^2\right] \\ & - \left(a\sum_{p=1}^{k} t_p + \left(\frac{1}{4} - b\right)k\right)\left[\sigma_{k+1}(z) + \frac{1}{2\alpha}\left\|x^{k+1} - x^k\right\|_2^2\right] \\ & + \frac{1}{2\alpha}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\left\|x^p - x^{p+1}\right\|_2^2 \\ & - \frac{1}{2\alpha}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\sum_{r=p}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|_2^2. \end{aligned} \tag{5.12}$$

Let us write the last two terms of the right-hand side for (5.12) as S_1 and S_2, respectively. (2.4) yields

$$\begin{aligned} S_2 &= -\frac{1}{2\alpha}\sum_{r=1}^{k}\sum_{p=1}^{r}\left(at_p - b + \frac{1}{4}\right)\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|_2^2 \\ &= -\frac{1}{2\alpha}\sum_{p=1}^{k}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right)\frac{1}{t_p}\left\|x^p - x^{p-1}\right\|_2^2. \end{aligned}$$

Hence, it follows that

$$S_1 + S_2 = -\frac{1}{2\alpha}\sum_{p=1}^{k}\left[\frac{1}{t_p}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right) - \left(at_p - b + \frac{1}{4}\right)\right]\left\|x^p - x^{p-1}\right\|_2^2$$
$$= -\frac{1}{2\alpha}\sum_{p=1}^{k}\frac{1}{t_p}\left[a\left(\sum_{r=1}^{p-1}t_r - t_p^2 + t_p\right) + \left(\frac{1}{4} - b\right)(p - t_p)\right]\left\|x^p - x^{p-1}\right\|_2^2. \quad (5.13)$$

Again $t_1 = 1$ gives

$$-t_p^2 + t_p = \sum_{r=1}^{p-1}(-t_{r+1}^2 + t_{r+1} + t_r^2 - t_r) = \sum_{r=1}^{p-1}\left(-(1-a)t_r - b + \frac{1}{4}\right)$$
$$= -(1-a)\sum_{r=1}^{p-1}t_r + \left(\frac{1}{4} - b\right)(p-1),$$

where the second equality comes from Lemma 5.1 (iii). Thus, we get

$$\sum_{r=1}^{p-1}t_r = \frac{t_p^2 - t_p}{1-a} + \left(\frac{1}{4} - b\right)\frac{p-1}{1-a}. \quad (5.14)$$

Substituting this into (5.13), it follows that

$$S_1 + S_2 = -\frac{1}{2\alpha(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|_2^2.$$

Combined with (5.12) and (5.14), we have

$$\begin{aligned} &t_{k+1}^2\sigma_{k+1}(z) - \sigma_1(z) \\ &\leq -\frac{1}{2\alpha}\left[\rho_k(z) - \left\|x^1 - z\right\|_2^2\right] \\ &\quad - \frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left[\sigma_{k+1}(z) + \frac{1}{2\alpha}\left\|x^{k+1} - x^k\right\|_2^2\right] \\ &\quad - \frac{1}{2\alpha(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|_2^2. \end{aligned}$$

Easy calculations give

$$\begin{aligned}\frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4}-b\right)k\right]\sigma_{k+1}(z) \\ + \frac{1}{2\alpha(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4}-b\right)k\right]\left\|x^{k+1}-x^k\right\|_2^2 \\ + \frac{1}{2\alpha(1-a)}\sum_{p=1}^{k}\left[a^2(t_p-1) + \left(\frac{1}{4}-b\right)\frac{p-t_p+a(t_p-1)}{t_p}\right]\left\|x^p-x^{p-1}\right\|_2^2 \\ + \frac{1}{2\alpha}\rho_k(z) \le \sigma_1(z) + \frac{1}{2\alpha}\left\|x^1-z\right\|_2^2.\end{aligned}$$

Lemma 5.2 (i) with $k=0$ and $y^1 = x^0$ and (2.3) with $(v^1, v^2, v^3) = (x^0, x^1, z)$ lead to

$$\sigma_1(z) \le -\frac{1}{2\alpha}\left[\left\|x^1-z\right\|_2^2 - \left\|x^0-z\right\|_2^2\right].$$

From the above two inequalities, we can derive the desired inequality. □

Let us define the linear function $P\colon \mathbf{R} \to \mathbf{R}$ and quadratic ones $Q_1\colon \mathbf{R} \to \mathbf{R}$, and $Q_2\colon \mathbf{R} \to \mathbf{R}$ by

$$\begin{aligned}P(\beta) &:= \frac{a^2(\beta-1)}{2}, \\ Q_1(\beta) &:= \frac{1-a}{4}\beta^2 + \left[1 - \frac{a}{2} + \frac{1-4b}{4(1-a)}\right]\beta + 1, \\ Q_2(\beta) &:= \frac{a(1-a)}{4}\beta^2 + \left[\frac{a}{2} + \frac{1-4b}{4(1-a)}\right]\beta.\end{aligned} \tag{5.15}$$

The following lemma provides the critical relation to evaluate the convergence rate of Algorithm 5.1.

Lemma 5.5

Under Assumption 4.4 with $\Omega \subseteq \mathbf{R}^n$, Algorithm 5.1 generates a sequence $\{x^k\}$ such that

$$Q_1(k)u_\infty\left(x^{k+1}\right) + \frac{1}{2\alpha}Q_2(k)\left\|x^{k+1}-x^k\right\|_2^2 + \frac{1}{2\alpha}\sum_{p=1}^{k}P(p)\left\|x^p-x^{p-1}\right\|_2^2 = O(1)$$

for all $k \ge 0$, where $O\colon [0,+\infty) \to \mathbf{R}$ satisfies $\limsup_{t\to\infty} O(t)/t < +\infty$, $P, Q_1, Q_2\colon \mathbf{R} \to \mathbf{R}$ are given in (5.15), respectively, and u_∞ is the gap function defined by (3.1).

Proof. Let $k \geq 0$. With similar arguments used in the proof of Theorem 4.4, we get

$$\sup_{z\in\Omega} \sigma_{k+1}(z) = u_\infty\big(x^{k+1}\big).$$

Since $\rho_k(z) \geq 0$ and Ω is bounded, Lemma 5.4 and the above equality lead to

$$\begin{aligned}\frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4}-b\right)k\right]\sigma_{k+1}(z) \\ + \frac{1}{2\alpha(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4}-b\right)k\right]\left\|x^{k+1}-x^k\right\|_2^2 \\ + \frac{1}{2\alpha(1-a)}\sum_{p=1}^{k}\left[a^2(t_p-1) + \left(\frac{1}{4}-b\right)\frac{p-t_p+a(t_p-1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|_2^2 \\ + \frac{1}{2\alpha}\rho_k(z) = O(1).\end{aligned}$$

We now show that the coefficients of the three terms on the right-hand side can be bounded from below by the polynomials given in (5.15). First, by using the relation

$$t_{k+1} \geq \frac{1-a}{2}k + 1 \tag{5.16}$$

obtained from Lemma 5.1 (i) and $a \in [0,1)$, we have

$$\begin{aligned}\frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4}-b\right)k\right] &= \frac{1}{1-a}\left[t_{k+1}(t_{k+1}-a) + \left(\frac{1}{4}-b\right)k\right] \\ &\geq \frac{1}{1-a}\left[\left(\frac{1-a}{2}k+1\right)\left(\frac{1-a}{2}k+1-a\right) + \left(\frac{1}{4}-b\right)k\right] = Q_1(k).\end{aligned}$$

Again, (5.16) gives

$$\begin{aligned}\frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4}-b\right)k\right] &= \frac{a}{1-a}t_{k+1}(t_{k+1}-1) + \frac{1-4b}{4(1-a)}k \\ &\geq \frac{a}{1-a}\left(\frac{1-a}{2}k+1\right)\left(\frac{1-a}{2}k\right) + \frac{1-4b}{4(1-a)}k = Q_2(k).\end{aligned}$$

Moreover, since $t_p \leq p$ (cf. Lemma 5.1 (ii)), $t_k \geq 1$ (cf. Lemma 5.1 (i)), and $b \in (a^2/4, 1/4]$, we obtain

$$\frac{1}{1-a}\left[a^2(t_p-1) + \left(\frac{1}{4}-b\right)\frac{p-t_p+a(t_p-1)}{t_p}\right] \geq \frac{a^2}{1-a}(t_p-1) \geq P(p).$$

Thus, combining the above equations, we get the desired inequality. □

Then, we can finally prove the main theorem.

Theorem 5.1 (ii). It is clear from Lemma 5.5 and $Q_1(k) = O(k^2)$ as $k \to \infty$. □

Remark 5.2
Lemma 5.5 also implies the following other claims than Theorem 5.1 (ii):

- $O(1/k^2)$ *convergence rate of* $\left\{ \left\| x^{k+1} - x^k \right\|_2^2 \right\}$ *when* $a > 0$*;*
- *the absolute convergence of* $\left\{ k \left\| x^{k+1} - x^k \right\|_2^2 \right\}$ *when* $a > 0$*;*

*Note that the second one generalizes [***Chambolle2015***] for single-objective problems.*

5.4 Convergence of the iterates

While the last section shows that Algorithm 5.1 has an $O(1/k^2)$ convergence rate, this section proves the following theorem:

Theorem 5.2
Let $\{x^k\}$ *be generated by Algorithm 5.1 with* $a > 0$*. Then, under Assumption 4.4, the following two properties hold:*

(i) $\{x^k\}$ *is bounded, and it has an accumulation point;*

(ii) $\{x^k\}$ *converges to a weak Pareto optimum for (1.1).*

The latter claim is also significant in the application. For example, finite-time manifold (active set) identification, which detects the low-dimensional manifold where the optimal solution belongs, essentially requires only the convergence of the generated sequence to a unique point rather than the strong convexity of the objective functions [**Sun2019**].

Again, we will prove Theorem 5.2 after showing some lemmas. The following lemma contributes strongly to the proof of the main theorem.

Lemma 5.6
Let $\{\gamma_q\}$ *be defined by line 9 in Algorithm 5.1. Then, we have*

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq 2(s-1) \quad \textit{for all } s, r \geq 1.$$

Proof. By using Lemma 5.1 (iv), we see that

$$\prod_{q=s}^{p} \gamma_q \leq \prod_{q=s}^{p} \frac{q-1}{q+1/2}.$$

Let Γ and B denote the gamma and beta functions defined by

$$\Gamma(\alpha) := \int_0^\infty \tau^{\alpha-1} \exp(-\tau)\, \mathrm{d}\tau \quad \text{and} \quad \mathrm{B}(\alpha,\beta) := \int_0^1 \tau^{\alpha-1}(1-\tau)^{\beta-1}\, \mathrm{d}\tau, \tag{5.17}$$

respectively. Applying the well-known properties:

$$\Gamma(\alpha) = (\alpha-1)!, \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \text{and} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{5.18}$$

we get

$$\prod_{q=s}^{p} \gamma_q \leq \frac{\Gamma(p)/\Gamma(s-1)}{\Gamma(p+3/2)/\Gamma(s+1/2)} = \frac{B(p,3/2)}{B(s-1,3/2)}.$$

This implies

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq \sum_{p=1}^{r} B(p,3/2)/B(s-1,3/2).$$

Then, it follows from the definition (5.17) of B that

$$\begin{aligned}
\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q &\leq \sum_{p=s}^{r} \int_0^1 \tau^{p-1}(1-\tau)^{1/2}\, \mathrm{d}\tau / B(s-1,3/2) \\
&= \int_0^1 \sum_{p=s}^{r} \tau^{p-1}(1-\tau)^{1/2}\, \mathrm{d}\tau / B(s-1,3/2) \\
&= \int_0^1 \frac{\tau^{s-1}-\tau^r}{1-\tau}(1-\tau)^{1/2}\, \mathrm{d}\tau / B(s-1,3/2) \\
&= \frac{B(s,1/2) - B(r+1,1/2)}{B(s-1,3/2)} \leq \frac{B(s,1/2)}{B(s-1,3/2)}.
\end{aligned}$$

Using again (5.18), we conclude that

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq \frac{\Gamma(s)\Gamma(1/2)/\Gamma(s+1/2)}{\Gamma(s-1)\Gamma(3/2)/\Gamma(s+1/2)} = 2(s-1).$$

□

Now, we introduce two functions $\omega_k \colon \mathbf{R}^n \to \mathbf{R}$ and $\nu_k \colon \mathbf{R}^n \to \mathbf{R}$ for any $k \geq 1$, which will help our analysis, by

$$\omega_k(z) \coloneqq \max\Big(0, \big\|x^k - z\big\|_2^2 - \big\|x^{k-1} - z\big\|_2^2\Big), \tag{5.19}$$

$$\nu_k(z) \coloneqq \big\|x^k - z\big\|_2^2 - \sum_{s=1}^{k} \omega_s(z). \tag{5.20}$$

The lemma below describes the properties of ω_k and ν_k.

Lemma 5.7

Let $\{x^k\}$ be generated by Algorithm 5.1 and recall that $\mathbf{lev}_F, \omega_k$, and ν_k are defined by (2.5), (5.19) and (5.20), respectively. Moreover, suppose that Assumption 4.4 holds with $\Omega \subseteq \mathbf{R}^n$ and that $\sigma_k(z) \geq 0$ for some $k \geq 1$ and $z \in \Omega$. Then, it follows for all $r = 1, \ldots, k$ that

(i) $\displaystyle\sum_{s=1}^{r} \omega_s(z) \leq \sum_{s=1}^{r} (6s-5)\big\|x^s - x^{s-1}\big\|_2^2$;

(ii) $\nu_{r+1}(z) \leq \nu_r(z)$.

Proof. Claim (i): Let $k \geq p \geq 1$. From the definition of y^{p+1} given in line 10 of Algorithm 5.1, we have

$$\begin{aligned}
&\big\|x^{p+1} - z\big\|_2^2 - \|x^p - z\|_2^2 \\
&= -\big\|x^{p+1} - x^p\big\|_2^2 + 2\big\langle x^{p+1} - y^{p+1}, x^{p+1} - z\big\rangle + 2\gamma_p\big\langle x^p - x^{p-1}, x^{p+1} - z\big\rangle \\
&= -\big\|x^{p+1} - x^p\big\|_2^2 + 2\big\langle x^{p+1} - y^{p+1}, y^{p+1} - z\big\rangle + 2\big\|x^{p+1} - y^{p+1}\big\|_2^2 \\
&\quad + 2\gamma_p\big\langle x^p - x^{p-1}, x^{p+1} - z\big\rangle.
\end{aligned}$$

On the other hand, Lemma 5.2 (i) gives

$$2\big\langle x^{p+1} - y^{p+1}, y^{p+1} - z\big\rangle \leq -2\alpha\sigma_{p+1}(z) - \big\|x^{p+1} - y^{p+1}\big\|_2^2.$$

Moreover, Lemma 5.3 with $(k_1, k_2) = (p+1, k+1)$ implies

$$\begin{aligned}
&-2\alpha\sigma_{p+1}(z) \\
&\quad \leq -\frac{2}{\ell}\sigma_{k+1}(z) - \big\|x^{k+1} - x^k\big\|_2^2 + \big\|x^{p+1} - x^p\big\|_2^2 - \sum_{r=p+1}^{k} \frac{1}{t_r}\big\|x^r - x^{r-1}\big\|_2^2 \\
&\quad \leq \big\|x^{p+1} - x^p\big\|_2^2,
\end{aligned}$$

where the second inequality comes from the assumption on $z \in \Omega$. Combining the above three inequalities, we get

$$\begin{aligned}\left\|x^{p+1}-z\right\|_2^2-\|x^p-z\|_2^2 &\leq \left\|x^{p+1}-y^{p+1}\right\|_2^2+2\gamma_p\left\langle x^p-x^{p-1},x^{p+1}-z\right\rangle \\ &= \left\|x^{p+1}-y^{p+1}\right\|_2^2+\gamma_p\Big(\|x^p-z\|_2^2-\left\|x^{p-1}-z\right\|_2^2+\left\|x^p-x^{p-1}\right\|_2^2 \\ &\qquad +2\left\langle x^p-x^{p-1},x^{p+1}-x^p\right\rangle\Big).\end{aligned}$$

Using the relation $\|x^{p+1}-y^{p+1}\|_2^2+2\gamma_p\langle x^p-x^{p-1},x^{p+1}-x^p\rangle = \|x^{p+1}-x^p\|_2^2+\gamma_p^2\|x^p-x^{p-1}\|_2^2$, which holds from the definition of y^k, we have

$$\begin{aligned}\left\|x^{p+1}-z\right\|_2^2-\|x^p-z\|_2^2 &\leq \left\|x^{p+1}-x^p\right\|_2^2 \\ &\quad +\gamma_p\Big(\|x^p-z\|_2^2-\left\|x^{p-1}-z\right\|_2^2\Big)+(\gamma_p+\gamma_p^2)\left\|x^p-x^{p-1}\right\|_2^2.\end{aligned}$$

Since $0 \leq \gamma_p \leq 1$ from Lemma 5.1 (iv), we obtain

$$\begin{aligned}&\left\|x^{p+1}-z\right\|_2^2-\|x^p-z\|_2^2 \\ &\quad\leq \gamma_p\Big(\|x^p-z\|_2^2-\left\|x^{p-1}-z\right\|_2^2+2\left\|x^p-x^{p-1}\right\|_2^2\Big)+\left\|x^{p+1}-x^p\right\|_2^2 \\ &\quad\leq \gamma_p\Big(\omega_p(z)+2\left\|x^p-x^{p-1}\right\|_2^2\Big)+\left\|x^{p+1}-x^p\right\|_2^2,\end{aligned}$$

where the second inequality follows from the definition (5.19) of ω_p. Since the right-hand side is non-negative, (5.19) again gives

$$\omega_{p+1}(z) \leq \gamma_p\Big(\omega_p(z)+2\left\|x^p-x^{p-1}\right\|_2^2\Big)+\left\|x^{p+1}-x^p\right\|_2^2.$$

Let $s \leq k$. Applying the above inequality recursively and using $\gamma_1 = 0$, we get

$$\begin{aligned}\omega_s(z) &\leq 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p-x^{p-1}\right\|_2^2+2\prod_{q=1}^{s}\gamma_q\left\|x^1-x^0\right\|_2^2+\left\|x^s-x^{s-1}\right\|_2^2 \\ &\leq 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p-x^{p-1}\right\|_2^2+\left\|x^s-x^{s-1}\right\|_2^2.\end{aligned}$$

Adding up the above inequality from $s = 1$ to $s = r \leq k$, we have

$$\begin{aligned}\sum_{s=1}^{r} \omega_s(z) &\leq 3 \sum_{s=1}^{r} \sum_{p=1}^{s} \prod_{q=p}^{s} \gamma_q \left\|x^p - x^{p-1}\right\|_2^2 + \sum_{s=1}^{r} \left\|x^s - x^{s-1}\right\|_2^2 \\ &= 3 \sum_{p=1}^{r} \sum_{s=p}^{r} \prod_{q=p}^{s} \gamma_q \left\|x^p - x^{p-1}\right\|_2^2 + \sum_{s=1}^{r} \left\|x^s - x^{s-1}\right\|_2^2 \\ &= \sum_{s=1}^{r} \left(3 \sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q + 1 \right) \left\|x^s - x^{s-1}\right\|_2^2,\end{aligned}$$

where the first equality follows from (2.4). Thus, Lemma 5.6 implies

$$\sum_{s=1}^{r} \omega_s(z) \leq \sum_{s=1}^{r} (6s - 5) \left\|x^s - x^{s-1}\right\|_2^2.$$

Claim (ii): (5.20) yields

$$\begin{aligned}\nu_{r+1}(z) &= \left\|x^{r+1} - z\right\|_2^2 - \omega_{r+1}(z) - \sum_{s=1}^{r} \omega_s(z) \\ &= \left\|x^{r+1} - z\right\|_2^2 - \max\left(0, \left\|x^{r+1} - z\right\|_2^2 - \left\|x^r - z\right\|_2^2\right) - \sum_{s=1}^{r} \omega_s(z) \\ &\leq \left\|x^{r+1} - z\right\|_2^2 - \left(\left\|x^{r+1} - z\right\|_2^2 - \left\|x^r - z\right\|_2^2\right) - \sum_{s=1}^{r} \omega_s(z) \\ &= \left\|x^r - z\right\|_2^2 - \sum_{s=1}^{r} \omega_s(z) = \nu_r(z),\end{aligned}$$

where the second and third equalities come from the definitions (5.19) and (5.20) of ω_{r+1} and ν_r, respectively. □

Let us now prove the first part of the main theorem.

Proof of Theorem 5.2 (i). Let $k \geq 1$ and suppose that $z \in \Omega$ satisfies $\sigma_k(z) \geq 0$. Then, Lemma 5.7 (ii) gives

$$\begin{aligned}\nu_k(z) \leq \nu_1(z) &= \left\|x^1 - z\right\|_2^2 - \omega_1(z) \\ &= \left\|x^1 - z\right\|_2^2 - \max\left(0, \left\|x^1 - z\right\|_2^2 - \left\|x^0 - z\right\|_2^2\right) \\ &\leq \left\|x^1 - z\right\|_2^2 - \left(\left\|x^1 - z\right\|_2^2 - \left\|x^0 - z\right\|_2^2\right) = \left\|x^0 - z\right\|_2^2,\end{aligned}$$

where the second equality follows from the definition (5.19) of ω_1. Considering the definition (5.20) of ν_k, we obtain

$$\left\|x^k - z\right\|_2^2 \leq \left\|x^0 - z\right\|_2^2 + \sum_{s=1}^{k} \omega_s(z).$$

Taking the square root of both sides and using (5.19), we get

$$\left\|x^k - z\right\| \leq \sqrt{\|x^0 - z\|_2^2 + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|_2^2}.$$

Applying the reverse triangle inequality $\left\|x^k - x^0\right\|_2 - \|x^0 - z\|_2 \leq \left\|x^k - z\right\|_2$ to the left-hand side leads to

$$\left\|x^k - x^0\right\| \leq \left\|x^0 - z\right\|_2 + \sqrt{\|x^0 - z\|_2^2 + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|_2^2}.$$

Since z belongs to a bounded set Ω and $a > 0$, the right-hand side is bounded from above according to Lemma 5.5. This implies that $\left\{x^k\right\}$ is bounded, and so it has accumulation points. □

Before proving Theorem 5.2 (ii), we show the following lemma.

Lemma 5.8
Let $\left\{x^k\right\}$ be generated by Algorithm 5.1 with $a > 0$ and suppose that Assumption 4.4 holds. Then, if $\bar{x}$ is an accumulation point of $\left\{x^k\right\}$, then $\left\{\left\|x^k - \bar{x}\right\|\right\}$ is convergent.

Proof. Assume that $\left\{x^{k_j}\right\} \subseteq \left\{x^k\right\}$ converges to $\bar{x}$. Then, we have $\sigma_{k_j}(\bar{x}) \to 0$ by the definition (5.11) of σ_{k_j}. Therefore, Lemma 5.7 with $z = \bar{x}$ and $k = \infty$ means that $\{\nu_k(\bar{x})\}$ is non-increasing and bounded, i.e., convergent. Hence $\left\{\left\|x^k - \bar{x}\right\|\right\}$ is convergent. □

Finally, we finish the proof of the main theorem.

Proof of Theorem 5.2 (ii). Suppose that $\left\{x^{k_j^1}\right\}$ and $\left\{x^{k_j^2}\right\}$ converges to $\bar{x}^1$ and $\bar{x}^2$, respectively. From Lemma 5.8, we see that

$$\lim_{j\to\infty}\left(\left\|x^{k_j^2} - \bar{x}^1\right\|_2^2 - \left\|x^{k_j^2} - \bar{x}_2^2\right\|^2\right) = \lim_{j\to\infty}\left(\left\|x^{k_j^1} - \bar{x}^1\right\|_2^2 - \left\|x^{k_j^1} - \bar{x}_2^2\right\|^2\right).$$

This yields that $\|\bar{x}^1 - \bar{x}^2\|_2^2 = -\|\bar{x}^1 - \bar{x}^2\|_2^2$, and so $\|\bar{x}^1 - \bar{x}^2\|_2^2 = 0$, i.e., $\{x^k\}$ is convergent. Let $x^k \to x^*$. Since $\left\|x^{k+1} - x^k\right\|_2^2 \to 0$, $\{y^k\}$ is also convergent to x^*. Therefore, Proposition 5.1 shows that x^* is weakly Pareto optimal for (1.1). □

5.5 Numerical experiments

This section compares the performance of Algorithm 5.1 with various a and b through numerical experiments. We run all experiments in Python 3.9.9 on a machine with 2.3 GHz Intel Core i7 CPU and 32 GB memory. For each example, we test 15 different hyperparameters combining $a = 0, 1/6, 1/4, 1/2, 3/4$ and $b = a^2/4, (a^2 + 1)/8, 1/4$, i.e.,

$$(a, b) = \left\{ \begin{array}{c} (0,0), (0,1/8), (0,1/4), \\ (1/6,1/144), (1/6,37/288), (1/6,1/4), \\ (1/4,1/64), (1/4,17/128), (1/4,1/4), \\ (1/2,1/16), (1/2,5/32), (1/2,1/4), \\ (3/4,9/64), (3/4,25/128), (3/4,1/4) \end{array} \right\},$$

and we set $\varepsilon = 10^{-5}$ for the stopping criteria.

5.5.1 Artificial test problems (bi-objective and tri-objective)

First, we solve the multi-objective test problems in the form (1.1), modifications from [**Jin2001**, **Fliege2009**], whose objective functions are defined by

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = g_2(x) = 0, \tag{JOS1}$$

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = \frac{1}{n}\|x\|_1, g_2(x) = \frac{1}{2n}\|x-1\|_1, \tag{JOS1-L1}$$

$$\begin{cases} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = 0, \end{cases} \tag{FDS}$$

$$\begin{cases} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = \delta_{\mathbf{R}^n_+}(x), \end{cases} \tag{FDS-CON}$$

where $x \in \mathbf{R}^n, n = 50$ and $\delta_{\mathbf{R}^n_+}$ is an indicator function (1.7) of the nonnegative orthant. We choose 1000 initial points, commonly for all pairs (a, b), and randomly with a uniform distribution between $\underline{c}$ and $\overline{c}$, where $\underline{c} = (-2, \ldots, -2)^\top$ and $\overline{c} = (4, \ldots, 4)^\top$ for (JOS1) and (JOS1-L1), $\underline{c} = (-2, \ldots, -2)^\top$ and $\overline{c} = (2, \ldots, 2)^\top$ for (FDS), and $\underline{c} = (0, \ldots, 0)^\top$ and $\overline{c} = (2, \ldots, 2)^\top$ for (FDS-CON). Moreover, we use backtracking for updating α to satisfy (5.10), with 1 as the initial value of α and 0.5 as the constant multiplied into α at each iteration. Furthermore, at each iteration, we transform the subproblem (5.1) into their dual like (3.16) and solve them with the trust-region interior point method [**Byrd1999**] using the scientific library SciPy.

Figure 5.1 and Table 5.1 present the experimental results. Figure 5.1 plots the solutions only for the cases $(a, b) = (0, 1/4), (3/4, 1/4)$, but other combinations also yield similar plots, including a wide range of Pareto solutions. Table 5.1 shows that the new momentum factors are fast enough to compete with the existing ones ($(a, b) = (0, 1/4)$ or $b = a^2/4$) and better than them in some cases.

Table 5.1: Average computational costs to solve the multi-objective examples

(a) (JOS1)

a	b	Time [s]	Iterations
0	0	6.442	97.0
0	1/8	5.158	81.217
0	1/4	4.207	65.0
1/6	1/144	4.244	67.0
1/6	37/288	5.182	82.0
1/6	1/4	4.268	66.0
1/4	1/64	6.224	99.0
1/4	17/128	7.239	113.566
1/4	1/4	3.205	51.0
1/2	1/16	4.51	72.0
1/2	5/32	4.562	71.0
1/2	1/4	4.466	70.0
3/4	9/64	4.323	67.998
3/4	25/128	3.104	49.0
3/4	1/4	3.741	47.0

(b) (JOS1-L1)

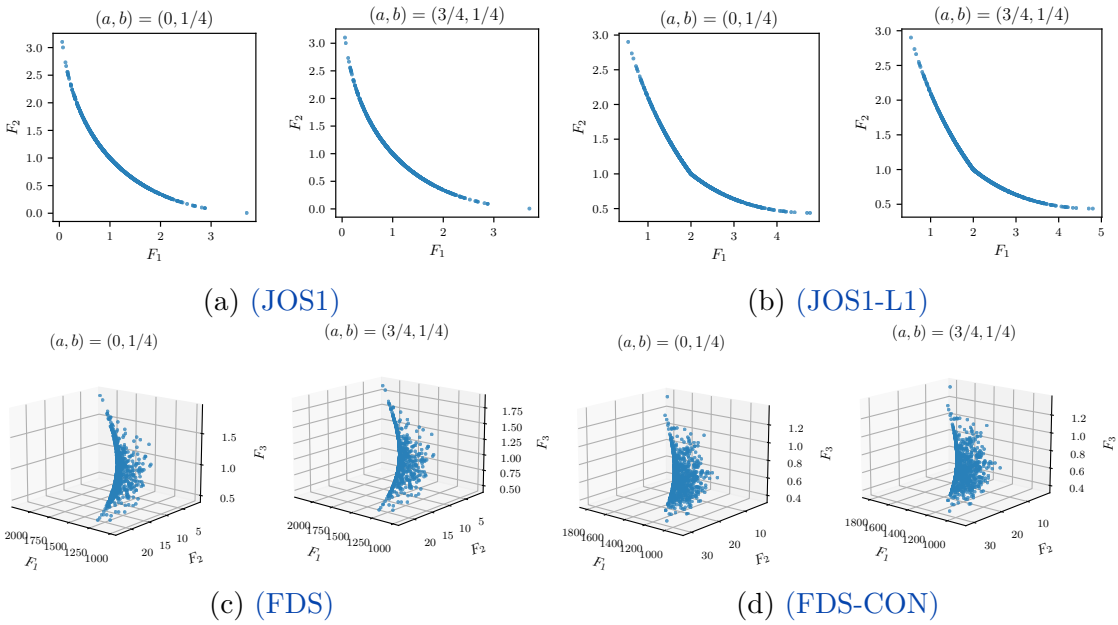
a	b	Time [s]	Iterations
0	0	10.733	157.512
0	1/8	11.054	161.065
0	1/4	11.122	161.734
1/6	1/144	9.85	141.731
1/6	37/288	9.994	144.863
1/6	1/4	10.399	150.592
1/4	1/64	9.271	135.804
1/4	17/128	9.463	137.108
1/4	1/4	9.662	139.848
1/2	1/16	7.439	109.082
1/2	5/32	7.642	110.204
1/2	1/4	7.723	111.599
3/4	9/64	5.253	77.366
3/4	25/128	5.39	79.425
3/4	1/4	5.678	82.37

(c) (FDS)

a	b	Time [s]	Iterations
0	0	29.24	204.438
0	1/8	29.797	210.595
0	1/4	30.565	214.934
1/6	1/144	24.964	174.393
1/6	37/288	25.375	177.944
1/6	1/4	26.065	182.398
1/4	1/64	22.94	159.737
1/4	17/128	23.311	162.629
1/4	1/4	23.976	166.918
1/2	1/16	17.909	122.653
1/2	5/32	18.14	123.96
1/2	1/4	18.221	125.697
3/4	9/64	13.584	94.176
3/4	25/128	13.674	94.705
3/4	1/4	13.795	94.868

(d) (FDS-CON)

a	b	Time [s]	Iterations
0	0	37.345	259.508
0	1/8	37.439	261.522
0	1/4	37.94	263.911
1/6	1/144	32.463	227.063
1/6	37/288	38.265	229.736
1/6	1/4	45.661	231.958
1/4	1/64	41.434	209.35
1/4	17/128	33.664	211.69
1/4	1/4	30.772	213.811
1/2	1/16	22.92	158.448
1/2	5/32	23.1	159.685
1/2	1/4	23.539	162.226
3/4	9/64	17.092	118.616
3/4	25/128	17.123	118.063
3/4	1/4	17.115	118.844



(a) (JOS1) (b) (JOS1-L1)

(c) (FDS) (d) (FDS-CON)

Figure 5.1: Pareto solutions obtained with some (a, b)

5.5.2 Image deblurring (single-objective)

Since our proposed momentum factor is also new in the single-objective context, we also tackle deblurring the cameraman test image via a single-objective ℓ_2-ℓ_1 minimization, inspired by [**Beck2009**]. In detail, as shown in Figure 5.2, to a 256×256 cameraman test image with each pixel scaled to $[0, 1]$, we generate an observed image by applying a Gaussian blur of size 9×9 and standard deviation 4 and adding a zero-mean white Gaussian noise with standard deviation 10^{-3}.

(a) Original (b) Blurred and noisy

Figure 5.2: Deblurring of the cameraman

Letting θ, B, and W be the observed image, the blur matrix, and the inverse of the Haar wavelet transform, respectively, consider the single-objective problem (1.1) with $m = 1$ and

$$f_1(x) \coloneqq \|BWx - \theta\|^2 \quad \text{and} \quad g_1(x) = \lambda\|x\|_1,$$

where $\lambda \coloneqq 2 \times 10^{-5}$ is a regularization parameter. Unlike in the previous subsection, we can compute ∇f's Lipschitz constant by calculating $(BW)^\top(BW)$'s eigenvalues using the two-dimensional cosine transform [**Hansen2006**], so we use it constantly as α^{-1}. Moreover, we use the observed image's Wavelet transform as the initial point.

Figure 5.3 shows the reconstructed image from the obtained solution. Images produced by all hyperparameters are similar, so we present only $(a, b) = (0, 1/4)$ and $(1/2, 1/4)$. Moreover, we summarize the numerical performance in Table 5.2 and Figure 5.4. Like the last subsection, this example suggests that our new momentum factors may occasionally improve the algorithm's performance even for single-objective problems.

(a) $(a, b) = (0, 1/4)$ (b) $(a, b) = (1/2, 1/4)$

Figure 5.3: Deblurred image

Table 5.2: Computational costs for the image deblurring

a	b	Total time [s]	Iteration counts
0	0	75.227	558
0	1/8	75.176	558
0	1/4	75.388	558
1/6	1/144	66.499	460
1/6	37/288	66.866	462
1/6	1/4	66.685	462
1/4	1/64	61.791	421
1/4	17/128	61.622	421
1/4	1/4	35.69	421
1/2	1/16	26.828	306
1/2	5/32	26.274	304
1/2	1/4	25.535	303
3/4	9/64	32.54	369
3/4	25/128	30.473	364
3/4	1/4	27.713	360

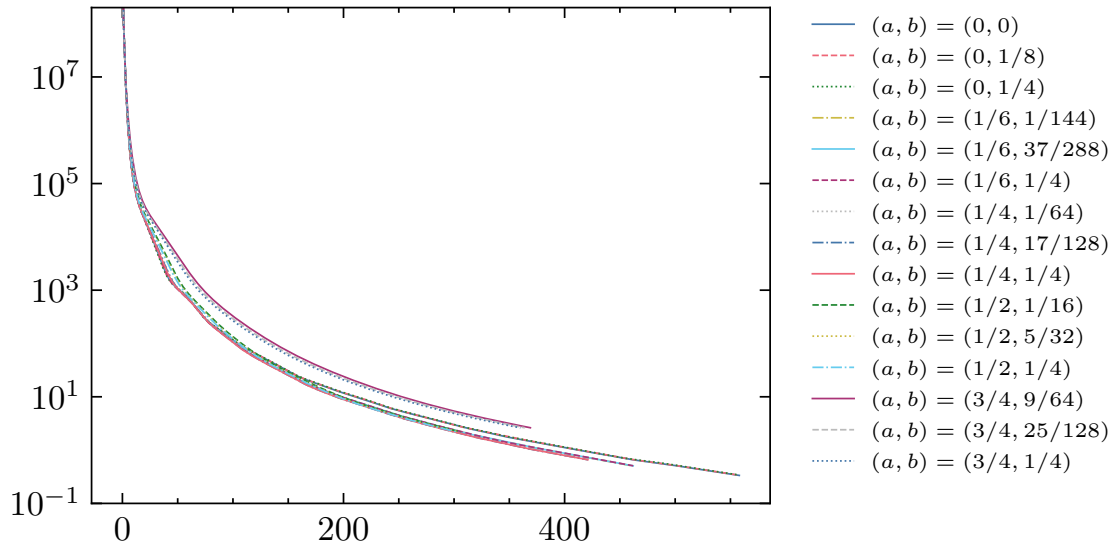


Figure 5.4: Values of $u_\infty\big(x^k\big) = F_1(x) - F_1(x^*)$, where x^* is the optimal solution estimated from the original image

5.6 Conclusions

We have successfully accelerated the proximal gradient method for multi-objective optimization by putting information on the previous points into the subproblem. Moreover, we have generalized the momentum factor in a form that is even new in the single-objective context and includes the known FISTA momentum factors [**Beck2009**, **Chambolle2015**]. Furthermore, with the proposed momentum factor, we proved under reasonable assumptions that the algorithm has an $O(1/k^2)$ convergence rate and that the iterates converge to Pareto solutions. Moreover, the numerical results reinforced these theoretical properties and suggested the potential for our new momentum factor to improve the performance.

Chapter 6

Conclusions

This thesis has proposed new merit functions, the proximal gradient method, and the accelerated proximal gradient method for non-smooth multi-objective optimization problems. We summarize the results obtained here as follows:

(i) In Chapter 3, we have proposed three merit functions for non-smooth multi-objective optimization: (a) the gap function for continuous multi-objective optimization; (b) the regularized gap function for convex multi-objective optimization; (c) the regularized and partially linearized gap function for composite multi-objective optimization. First, we have shown that they satisfy the properties as merit functions and proved the lower semi-continuity of the item (a) and the locally Lipschitz continuity of the items (b) and (c). We have also confirmed the differentiability of the items (b) and (c) under reasonable assumptions and that the stationary points of the items (b) and (c) solve the original multi-objective problem under strict convexity. Secondly, we have derived inequalities among different merit functions under certain conditions. We thirdly have demonstrated that the level-boundedness of the objective functions implies the level-boundedness of the associated merit functions. Finally, we proposed the multi-objective proximal-PL condition, weaker than the strong convexity, and proved that it provides the error-bound property of the proposed merit functions.

(ii) In Chapter 4, we have developed the proximal gradient method for composite multi-objective optimization. We have shown that every accumulation point of the generated sequence, if it exists, is Pareto stationary. Moreover, we presented global convergence rates for the proposed algorithm, matching what

we know in scalar optimization for non-convex $O(\sqrt{1/k})$ and convex $O(1/k)$ ones. We also have extended the so-called Polyak-Łojasieqicz (PL) inequality for multi-objective optimization and established the linear convergence rate for multi-objective problems that satisfy such inequalities. Furthermore, we have converted the subproblems to well-known convex optimization problems for the robust multi-objective problem. Finally, we have reported some numerical results.

(iii) In Chapter 5, we have proposed the accelerated proximal gradient method for convex composite multi-objective optimization. We have proved the proposed methods' $O(1/k^2)$ convergence rate and the global convergence property. This method includes some hyperparameters, which is new even for single-objective cases. We finally have reported some numerical results, showing that some of these choices give better results than the classical algorithms.

We believe that these contributions have had some impact on non-smooth and composite multi-objective optimization. However, there are still many open problems. We conclude this thesis by describing future works related to our results.

(i) We can consider our proposed merit function's natural extension to infinite-dimensional vector optimization. We can also regard other famous merit functions' generalization to multi-objective or vector problems, such as the implicit Lagrangian [**Mangasarian1993**] and the squared Fischer-Burmeister function [**Kanzow1996**]. Moreover, developing a new multi-objective algorithm using such merit functions would be interesting.

(ii) Extending the many variants of the proximal gradient method in single-objective optimization to multi-objective optimization problems is a challenge that needs addressing. Obtaining a theoretically sound extension will not be straightforward for any method. However, we believe that finding practical applications of composite multi-objective optimization, such as machine learning, will significantly impact this field.

(iii) Our proposed method has the potential to achieve finite-time manifold (active set) identification [**Sun2019**] without the assumption of the strong convexity (or its generalizations such as PL conditions or error bounds [**Karimi2016**]). Moreover, we took a single update rule of t_k for all iterations in this work, but the adaptive change of the strategy in each iteration is conceivable, which

has the potential to guarantee linear convergence under PL conditions, as in [**Aujol2021**]. It might also be interesting to estimate the Lipschitz constant simultaneously with that change, like in [**Scheinberg2014**]. In addition, an extension to the inexact scheme like [**Villa2013**] would be significant. Furthermore, it is crucial to extend the variants of the accelerated proximal gradient method to multi-objective optimization. Moreover, applying our acceleration techniques to large-scale problems like stochastic accelerated gradient descent would be interesting. Developing internal techniques, such as a warm start for subproblems and inexact methods, would also be necessary for applications.